இயற்கை மொழி ஆய்வு

பேராசிரியர் ச. இராசேந்திரன்
அமிர்தா விஷ்வ வித்யபீடம்
கோயம்பத்தூர்

rajushush@gamil.com, 2005

பொருளடக்கம்

பக்கம்

இயல் 1: அறிமுகம்

1.4. இயற்கை மொழி ஆய்வு

1.4.1.இற்கை மொழியும், செயற்கை மொழியும்

1.4.2.பேச்சைத் தெரிந்து கொள்ளல்

1.4.3 இயந்திர மொழிபெயர்ப்பு

1.4.4. கணிப்பொறி வழி மொழிப்பகுப்பாய்வு

1.5. கணிப்பொறி வழி மொழி ஆய்வின் தேவைகளும் சிக்கல்களும்

1.6. கணிப்பொறியில் மொழிகளைக் கையாளுதல் - வகைகள்

1.7. மொழி ஆய்விற்கு தேவைப்படும் கருவிகள்

1.8. மொழியாய்வில் மொழியியலாளர்களுக்குத் தேவைப்படும் திறமைகள்

1.9. மொழியாய்வில் கணிப்பொறியாளர்க்குத் தேவைப்படும் திறமைகள்

1.10. தமிழில் இயற்கை மொழி ஆய்வு

1.10.1. தமிழில் இயற்கை மொழி ஆய்வுகான முன்னேற்பாடுகள்

1.10.2. மொழித்தொழில் நுட்பம்

1.10.3. தமிழ்மொழித் தொழில் நுட்ப வளர்ச்சியின் இன்றைய நிலை

இயல் 2

கணிப்பொறி முலம் தமிழ்மொழி ஆய்வு

2.1. மொழி ஆய்வும் கணிப்பொறியும்

2.2. உரையிலிருந்து பேச்சு

2.2.1. பேச்சு ஒலிகளும் ஒலி பெயர்ப்பும்

2.2.2. ஒலியன்களும் ஒலியன் விதிகளும்

2.2.3. ஒலியனியல் விதிகளும் முற்றுநிலை மாற்றிகளும்

2.2.4. உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல்

2.2.4.1. உச்சரிப்பு அகராதிகள்

2.2.4.2. அகராதியை நோக்குவதற்கு அப்பாற்பட்ட உரை ஆய்வு

2.2.4.3. முற்றுநிலைமாற்றி அடிப்படையில் உச்சரிப்பு அகராதி

2.2.5. உரையிலிருந்து-பேச்சில் மீக்கூறு

2.2.5.1. மீக்கூறின் ஒலியல் அல்லது ஒலியியக்கவியல் நோக்குகள்

2.2.5.2. பேச்சு உருவாக்கத்தில் மீக்கூறு

2.2.6. தமிழில் உரையிலிருந்து பேச்சு

2.3. எழுத்துத்திருத்தியும் இலக்கணப்பிழைத் திருத்தியும்

2.3.1. எழுத்துப் பிழை நிகழ்வதற்கான வாய்ப்புகள்

2.3.2. பிழை நீக்கும் வழிமுறைகள்

2.3.3. எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் சிக்கல்களின் வகைகள்

2.3.4. மொழியில் இல்லாதச் சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல்திட்டம்

2.3.5. எழுத்துப்பிழைத் திருத்தி (Bayesian Medhod)

2.4. உருபனியல் ஆய்வு

2.4.1. உருபனியல் ஆய்விகளும் கோட்பாட்டு அணுகுமுறையும்

2.4.2. தமிழில் உருபனியல் ஆய்விகள்

2.4.2.1. இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி

2.4.2.2. கணேசனின் உருபனியல் பகுப்பாய்வி

2.4.2.3. கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி

2.4.2.4. தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்

2.4.2.5. AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி

2.4.2.6. வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்

2.4.2.7. வின்ஸ்டன் குருசின் ஜி.எஸ். மார்.்ப் பகுப்பாய்வியும் உருவாக்கியும்

2.4.2.8. துரைபாண்டியின் உருபனியல் பகுப்பாய்வு

2.4.2.9. RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி

2.4.2.10. RCILTS-T –இன் உருபனியல் உருவாக்கி

2.4.2.11. பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்

2.5. கணினி தொடரியல்

2.5.1. கணிப்பெறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்

2.5.2. ஆழமில்லாப் பகுப்பாய்வு

2.5.3. தமிழில் சொல்வகை அடையாளப்படுத்திகள்

2.5.4. தமிழில் தொடரியல் பகுத்துக் குறிப்பான்கள்

2.5.4.1. பாஸ்கரனின் தொடரியல் பகுத்துக்குறிப்பான்

2.5.4.2. குமார சண்முகத்தின் தமிழ்த்தொடரியல் பகுத்துக்குறிப்பான்

2.6. தமிழில் மொழிநடை ஆய்வி

2.7. கணினிப் பொருண்மையியல்

2.7.1. பொருண்மையை உருப்படுத்தம் செய்தல்

2.7.2. பொருண்மையியல் ஆய்வு

2.7.3. தொடரியலால் இயக்கப்படும் பொருண்மையியல்

2.7.4. சூழல் கட்டுப்பாடில்லாத விதிகளில் பொருண்மையின் விரிவாக்கங்கள்

2.7.5. முந்தையப்பகுத்துக்குறிப்பானில் பொருண்மையியல் ஆய்வை ஒருங்கிணைத்தல்

2.7.6. சொற்பொருண்மையியல்

2.7.7. தமிழில் கணினி பொருண்மையில் ஆய்வு

2.8. தமிழில் சொற்களஞ்சியம் தயாரித்தல்

2.9. தமிழில் சொல்வலை தயாரித்தல்

2.10. தமிழில் சொற்களை அகரவரிசைப் படுத்துதல்

2.11. தமிழில் சொல்லடைவு தயாரித்தல்

2.12. தமிழில் சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவுத் தயாரித்தல்

2.13. தமிழ்ச் சொல்லாய்வி தயாரித்தல்

2.14. எழுத்துப் பெயர்ப்பு செய்தல்

2.15. எழுத்துணரி உருவாக்குதல்

2.16. தமிழில் கணிப்பொறி வழி அகராதி தயாரித்தல்

2.16.1. அகராதி உருவாக்கும் படிநிலைகள்

2.16.2. கணிப்பொறி வழி அகராதி: வரலாற்றுச் சுருக்கம்

2.16.3. கணிப்பொறி வழி அகராதி: முறைமைகள்

2.16.4. அகராதி தயாரித்தலில் இயற்கை மொழியாய்வு

2.16.5. முடிவுரை

2.17. தமிழில் கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல்

2.17.1. கற்றல், கற்பித்தலில் கணிப்பொறி

2.17.2. கணிப்பொறி வழி கற்றல்-கற்பித்தலின் சிறப்பியல்புகள்

2.17.3. CAI- உருவாக்குவதற்கான கணிப்பொறி மொழிகள்

2.17.4. CAI உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள்

2.17.5. CAI–இல் நகரும் படங்கள்

2.17.6. CAI உருவாக்கத்தில் மனித நுண்ணறிவின் பங்கு

2.17.7.கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்

2.17.8.கணிப்பொறி வழி தமிழ் கற்றல்-கற்பித்தலுக்கான மென்பொருள்கள்

2.17.9. கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு

2.17.10. மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு

2.18. கணிப்பொறி வழி மொழிபெயர்ப்பு

2.18.1. கணிப்பொறி மொழி பெயர்ப்பின் வகைகள்

2.18.2. கணிப்பொறி வழி மொழி பெயர்ப்பின் வரலாறு

2.18.3. கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

2.18.4. உருஷ்யன்-தமிழ் மொழி பெயர்ப்பு ஒரு கண்ணோட்டம்

2.18.4.1. முன்னேற்றப் பணிகள்

2.18.4.2 உருசியன்-தமிழ் மொழி பெயர்ப்பின் படிநிலைகள்

2.18.5. கணிப்பொறியின் மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

2.18.5.1. உருபனியல் சிக்கல்

2.18.5.2. தொடரமைப்புச் சிக்கல்

2.18.5.3. சொல்வகைப்பாடு பொருத்தச் சிக்கல்

2.18.5.4. கருத்தாடல் சிக்கல்

2.18.5.5. கலாச்சார வேறுபாட்டுச் சிக்கல்

2.18.5.6. தொடரியல் பகுப்பாய்வுச் சிக்கல்

2.18.5.7. அகராதிச் சிக்கல்

2.18.5.8. பிற சிக்கல்கள்

2.18.6. தமிழில் கணிப்பொறி வழி மொழிபெயர்புக்கான ஆயத்தங்கள்

2.18.6.1. தமிழ் சார் இயந்திர மொழிபெயப்பு முயற்சிகள்

2.18.6.1.1 தமிழ் உருஷ்யன் மொழிப் பெயர்ப்புத்திட்டம்

2.18.6.1.2 உலக வலைப்பின்னல் மொழி – தமிழ்க்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

2.18.6.1.3 ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

2.18.6.2 தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்

2.18.6.2.1 உருபனியல் ஆய்வு **(Morphological Analysis)**

2.18.6.2.2 சொல் வகைப்பாட்டு அடையாளப்படுத்தல் **(Parts of Speech Tagging)**

2.18.6.2.3 தொடரியல் ஆய்வு **(Syntax analysis)**

2.18.6.2.4 பொருண்மையியல் ஆய்வு **(Semantic Analysis)**

2.18.6.2.5 தொடரியல் சார் மாற்றம் **(Syntactic Transfer)**

2 .18.6.2.6 அகராதி பொருத்தம் (Dictionary Mapping)

2.18.6.2.7 உருபனியல் உருவாக்கம் (Morphdogical Generation)

2.18.6.3 தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்

2.8.3.1 கருவிகள் (Tools)

2.18.6.3.1.1 உருபனியல் ஆய்வி (Morphoogical Analyzer)

2.18.6.3.1.2 அடையாளப்படுத்தி

2.18.6.3.1.3 பகுத்துக் குறிப்பான் (Parser)

2.18.6.3.1.5 பொருண்மையியல் ஆய்வி (Semantic Analyser)

2.18.6.3.1.6 உருபனியல் உருவாக்கி (Morphological Generator)

2.18.6.3.2 மூலவளங்கள் அகராதி (Resources)

2.18.6.3.2.1 அகராதி (Dictionary)

2.18.6.3.2.2 மொழி கடந்த அகராதி (Cross -lingual Dictionary)

2.18.6.3.2.3 தரவுத்தொகுதி (Corpus)

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

இணை தரவுத்தொகுதிகள் (Parellel Corpora)

2.18.6.3.2.4 மாற்றமைவு இலக்கணம்

2.18.6.3.2.5 இணைவமைதி அகராதி

2.18.6.9 நிறைவுரை

இயல் 3

முடிவுரை

துணைநூல்கள்

இயல் 1
அறிமுகம்

1.1. நூலின் நோக்கம்

இன்றைய சூழ்நிலையின் அடிப்படையில் "இயற்கை மொழி ஆய்வு" என்ற தலைப்பில் அமைந்த இந்நூலில் தமிழில் இயற்கை மொழி ஆய்வு குறித்து இதுவரை வெளிவந்துள்ள ஆய்வேடுகள், கட்டுரைகள், நூல்கள் மற்றும் இணைய தளங்களில் உள்ள தகவல்கள் இவற்றின் அடிப்படையில் செய்திகளைத் திரட்டித்தர முயற்சி எடுக்கப்பட்டள்ளது.

1.2. நூலின் பயன்பாடு

இது ஒரு தொகுப்பு நூல் என்றாலும் இத்தகைய ஆய்வு அடுத்த கட்ட ஆய்வுகளுக்குப் பயனுள்ளதாய் அமையும். தமிழில் இயற்கை மொழி குறித்த பல ஆய்வுகள் மேற்கொள்ளப்பட்டாலும் வளந்து வரும் இவ்வாய்வுக்களத்தில் ஏற்பட்டுள்ள முன்னேற்றம் குறித்து மாணவர் சமுதாயம் அறியவேண்டும் அவர்கள் இதன் அடிப்படையில் அடுத்த கட்ட உயர்நிலை ஆய்வுகளை மேற்கொள்ள வேண்டும் என்ற ஆர்வத்தைப் தூண்டும் விதமாக இந்நூல் அமைந்துள்ளது. தமிழில் வெளிவந்துள்ள இயற்கை மொழி ஆய்வுகளின் நிறைகுறைகளை ஆய்ந்து மதிப்பீடு செய்யும் ஒரு ஆய்வு மிகவும் தேவைப்பட்டாலும் அத்தகைய முயற்சி இங்கு எடுக்கப்படவில்லை.

1.3. நூலில் அமைப்பு

இந்நூல் நான்கு பகுதிகளாகப் பிரிக்கப்பட்டுள்ளது: அறிமுகம், கணிப்பொறி வழி மொழிப்பகுப்பாய்வுகள், கணிப்பொறி வழி மொழிக் கருவிகள் தயாரித்தல், இறுதியுரை. அறிமுகத்தில் இயற்கை மொழி ஆய்வு குறித்தும் தமிழில் இயற்கை மொழி ஆய்வுகுறித்தும், குறிப்பாகத் தமிழ் தொழிநுட்ப வளர்ச்சியின் இன்றைய நிலை குறித்தும், கணிப்பொறி வழி மொழியியல் ஆய்வின் தேவைகள் குறித்தும், கணிப்பொறியில் மொழிகளைக் கையாளுதல் குறித்தும், மொழி ஆய்விற்குத் தேவைப்படும் கருவிகள் குறித்தும், மொழி ஆய்வில் கணிப்பொறியாளருக்கும் மொழியியலாளருக்கும் தேவையான திறமைகள் குறித்தும் மொழியியல் விளக்கப்பட்டுள்ளன. மேலும் இவ்வாய்வின் நோக்கம், பயன்பாடு, ஆய்வேட்டின் அமைப்பு என்பனவும் அறிமுகத்தில் தரப்பட்டுள்ளன. இரண்டாவது பகுதியில் ஒலியனியல், உருபனியல், தொடரியல், பொருண்மையியல் ஆகிய மட்டங்களில் கணிப்பொறி வழி தமிழுக்குச் செய்யப்பட்டுள்ள அல்லது முயற்சிக்கப்பட்டுள்ள ஆய்வுகள் குறித்துச் சுருக்கமாகத் தரப்பட்டுள்ளன. மூன்றாவது பகுதியல் தமிழுக்கு மின்னகராதி, மின்சொற்களஞ்சியம், சொல்லாய்வி, கற்றல் கற்பித்தல் கருவி, எழுத்துத்திருத்தி, இலக்கண திருத்தி, இயந்திர மொழிபெயர்ப்புக் கருவி போன்றன உருவாக்குதல் குறித்தும் இதற்கான முயற்சிகள் குறித்தும் விரிவாக விளக்கப்பட்டுள்ளன. நான்காவது பகுதி இறுதியுரையாகும்.

1.4. இயற்கை மொழி ஆய்வு

கணிப்பொறி ஒரு மொழியாய்வுக் கருவி. செய்திகளை 1,0 என்ற எண்களால் ஆய்வது தான் கணிப்பொறி. மனித மூளை 0,1 என்ற அடிப்படை எண்களை வைத்துக் கொண்டு எந்திர மொழி மூலம் சிக்கலான கணிப்பொறியை உருவாக்கியது. எழுத்துக்களும், எண்களும் தனித்தன்மையான குறியீடுகளும் (special symbals) 0,1 என்பவைகளால் தான் உருப்படுத்தம் செய்யப்படுகின்றன. உயர்நிலை மொழிகள்

பொருள்கோளின் (interperters) ஒன்று கூட்டிகள் (assemblers) என்பன மொழியியலால் பயன்பெற்றன. இப்பொழுது மொழியியலார் கணினியை அவர்களின் கடினமான செயல்பாடுகளைச் செய்யும் உண்மையான கருவியாகப் பயன்படுத்தும் காலம் வந்துவிட்டது. மொழியியல் சார் கணிப்பொறி ஆய்வை கணினி மொழியியல் ஆய்வு அல்லது இயற்கை மொழி ஆய்வு என்பர். இது பேச்சைத் தெரிந்துகொள்ளுதல், பேச்சை உருவாக்குதல், உரைகளை/உரைகளை பேசாக்குதல், பேச்சுக்களைப் உரை ஆக்குதல், இயந்திர மொழிபெயர்ப்பு, விரிதரவுகளை ஆய்தல், உரைகளை ஆய்தல், மொழி அலகுகளைக் பகுப்பாய்வு செய்தல் என்பனவற்றை உள்ளடக்கும். மொழியாய்வு மொழியைப் புரிந்துக் கொள்ளுதல் என்பதுடன் பின்வருவனவற்றையும் உள்ளடக்கும். உரை அல்லது பேச்சுச் செய்திகளை பொருள் கொண்டுப் புரிந்துக் கொள்ளுதல். மொழிபெயர்த்தல், உரைகள் அல்லது பேச்சுச் செய்திகளை உருவாக்குதல்.

1.4.1. இயற்கை மொழியும் செயற்கைமொழியும்

மனிதர்கள் இயற்கையாகப் பயன்படுத்தும் எந்த மொழியும் இயற்கை மொழியாகும். செயற்கையாக மனிதனால் உருவாக்கப்பட்ட மொழிகள் இயற்கை மொழியில் இருந்து வேறுபட்டவை. எடுத்துக்காட்டாக, கணிப்பொறியில் பயன்படுத்தும் (programming) மொழிகளான போர்ட்ரான் (Fortran) பேசிக் (Basic), சி (C), புரோலாக் (Prolog) லிஸ்ப், (Lisp) பெர்ல் (Perl) என்பவைகளெல்லாம் செயற்கை மொழிகளாகும்.

இயற்கை மொழியாய்வு (natural language processing) கணிப்பொறியைக் கொண்டு மொழியை ஆய்வு செய்யும் எல்லா நடவடிக்கைகளையும் உள்ளடக்கும். பேச்சையும் (speech) எழுத்து மொழியையும் ஆயும் கணினி நுட்பங்கள் இதில் அடங்கும். பேச்சு உருவாக்கம் (speech synthesis) இயற்கை ஒலியான பேச்சை உருவாக்குவது நுட்ப

அடிப்படையில் சிக்கலானதாகும். என்ன பேசப்படுகிறது என்பதைப் புரிந்து கொள்ள சரியான இசையோட்டம் போன்றவை தேவை.

1.4.2. பேச்சைத் தெரிந்து கொள்ளல்

இது அடிப்படையில் தொடர்ச்சியான ஒலி அலைகளை வேறுபட்ட சொற்களாகப் பகுத்தல், இயற்கைமொழியைப் புரிந்துக் கொள்ளுதல் (natural language understanding), தனித்தனிச் சொற்களிலிருந்து பொருண்மைக்குச் செல்லுதல், இயற்கை மொழியை உருவாக்குதல், (natural language generation), அனுமானிக்க இயலாத உள்ளீடுகளுக்குப் பொருத்தமான பதில்களை உருவாக்குதல் என்பனவற்றை உள்ளடக்கும்.

1.4.3 இயந்திர மொழிபெயர்ப்பு

இது ஒரு இயற்கை மொழியை இன்னொன்றாக மொழிப்பெயர்ப்பு செய்தலைக் குறிப்பிடும். பொதுவாக இயற்கை மொழியாய்வு பேச்சை உட்படுத்தாமல் தான் பயன்படுத்தப்படுகின்றது. பேச்சு மற்றும் இயற்கை மொழியாய்வு என்பதைப் பேச்சையும் பிற இயற்கை மொழியாய்வுகளையும் உள்ளடக்கப் பயன்படுத்தலாம்.

1.4.4. கணிப்பொறி வழி மொழிப்பகுப்பாய்வு

சிக்கலான மொழி பழக்கத்தில் ஈடுப்பட தேவையானவைகளை ஆறு வேறுப்பட்ட பாகுபாடுகளாகப் பகுக்கலாம். ஒலியனியல் பகுப்பாய்வு, உருபனியல் பகுப்பாய்வு, தொடரியல் பகுப்பாய்வு, பொருண்மையியல் பகுப்பாய்வு, பயன்வழியியல் பகுப்பாய்வு, கருத்தாடல் பகுப்பாய்வு என்பர். மொழியை இம்மட்டங்களில் பகுப்பதை

கணிப்பொறி வழி மொழிப்பகுப்பாய்வு என்பர். இவை கணினி மொழியியலுக்கு அடிப்படை ஆய்வுகளாகும்.

ஒலியனியல் பகுப்பாய்வு: ஒலியனியல் பகுப்பாய்வு ஒலி அடையாள குறிகளிடமிருந்து அல்லது கேட்பு அடையாளக் குறிகளிடமிருந்து சொற்களைப் பிரித்தெடுக்கும், பேச்சைத் தெரிந்து கொள்ளும்.

ஒலிகள் → சொற்கள்
←
/b/+/ɔ:/ +t → bɔ:t/ ‘boat’
←

க்+ஆ+க்+அ+ம் → காலம்
←

உருபனியல் பகுப்பாய்வு: இது உருபன்களிலிருந்து சொற்களைப் பெறல் சொற்களை உருபன்பகளாகப் பகுத்தல் என்பனவற்றை உள்ளடக்கும்.

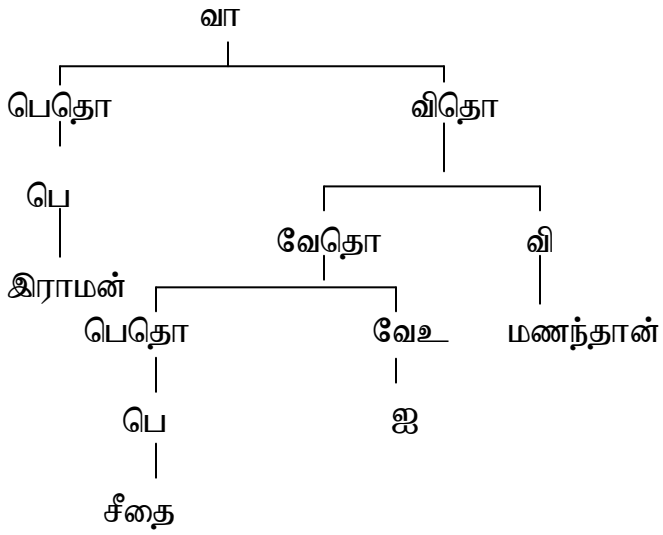
உருபன்கள் → சொற்கள்
←

மரம்+கள் → மரங்கள்
←

தொடரியல் பகுப்பாய்வு: இது சொற்களின் வரிசை அமைப்பிலிருந்து வாக்கிய அமைப்பைப்பெறல் அல்லது வாக்கியங்களைச் சொற்களாகப் பிரித்தல் என்பனவற்றை உள்ளடக்கும்.

சொல்வரிசை → வாக்கிய அமைப்பு
←

'இராமன் சீதையை மணந்தான்' என்ற சொல் வரிசையிலிருந்து பின் வரும் பகுத்தாயப் பட்ட கிளை அமைப்பைப் பெறலாம்.



பொருண்மையியல் பகுப்பாய்வு: வாக்கிய அமைப்பு சொற்பொருண்மை இவற்றிலிருந்து வாக்கியப் பொருண்மையைப் பெறல். எடுத்துக்காட்டாக, *இராமன் சீதையை மணந்தான்* என்று பகுத்தாயப்பட்ட கிளை அமைப்பிலிருந்து பின்வரும் தர்க்கக் கூற்றைப் பெறலாம்: மணந்தான் (இராமன், சீதை).

பயன்மொழியியல் பகுப்பாய்வு: வாக்கியப்பொருள் மற்றும் சூழல் இவற்றிலிருந்து சரியானப் பொருளைப்பெறல். இது பொது அறிவையும் முன்னர் வந்த கூற்றின் அறிவையும் உள்ளடக்கும். *அவன் சீதையை மணந்தான்* என்ற வாக்கியத்தில் வரும் பதிலீடுப் பெயர் (அவன்) முன்னர் வந்த வாக்கியத்தின் அறிவால் அவன் = ராமன் என்று புரிந்து கொள்ளப்படும்.

1.5. கணிப்பொறி வழி மொழி ஆய்வின் தேவைகளும் சிக்கல்களும்

கணிப்பொறியில் ஆங்கிலத்தைத் தவிர பிற மொழிகளையும் கையாளும் சூழ்நிலை வளர்ந்துள்ளது. குறிப்பாகக் கணிப்பொறியைத் தமிழில் கையாளுதல் கடந்த பத்து ஆய்டுகளுக்கு முன் தொடங்கப் பெற்று வளர்ச்சி அடைந்து வருகிறது. பொதுவான பயன்பாடுகளுக்குத் தமிழைக் கையாளுதல் கணிப்பொறியில் சமூக மேம்பாட்டை எய்த உதவுகிறது. இது தவிர மொழி ஆய்வுப் பணிகளுக்காகவும் கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

1.6. கணிப்பொறியில் மொழிகளைக் கையாளுதல் - வகைகள்

கணிப்பொறியில் மொழிகள் இரு விதமாகக் கையாளப் படுகின்றன:

1. அன்றாட பணிகளுக்காகக் கணிப்பொறியில் மொழியைக் கையாளுதல்
2. மொழி ஆய்வுக்காகக் கணிப்பொறியைக் கையாளுதல்

முதல் வகைப் பயன்பாட்டில் கணிப்பொறியில் மொழிகள் கையாளுவது கீழ்கண்ட பணிகளுக்கு உதவுகின்றன:

1. தமிழ்ச் சொல்லாய்விகளை உருவாக்குவதன் மூலம் கணிப்பொறியை ஒரு உயர்ரக மின்மத் தட்டச்சு போன்று தமிழுக்குப் பயன்படுத்துதல்
2. டி.டி.பி அதாவது மேஜைமேல் அச்சிடுதல் (Desk Top Printing (DTP)) போன்றவைகளுக்கான மென்பொருள்களைத் தமிழுக்கு உருவாக்கி அச்சுப் பணிகளுக்குக் கணிப்பொறியைப் பயன்படுத்தல்
3. நிறுவனங்களில் பணியாளர்கள் பட்டியல்கள், ஊதியப்பட்டியல்கள் மற்றும் வரவு செலவு அறிக்கைகளைத் தமிழில் தயாரித்தல்

4. வணிக நிறுவனங்களில் பொருள் வாங்கும் மற்றும் விற்கும் பட்டியல்களைக் கணிப்பொறி வழி தமிழில் தயாரித்தல்
5. தொலைவரி, பேஜர் போன்ற கருவிகளில் செய்தி தொடர்புக்குத் தமிழைக் கையாளுதல்

மேற்கண்ட பணிகளுக்காகக் கணிப்பொறியைக் கையயாளுவது தொடங்கப்பட்டுள்ளது. இத்தகைய வசதிகள் பொதுமக்களுக்குப் பயன்தரத்தக்க வகையில் அமைந்துள்ளன. எனவே இத்தகைய மென்பொருள்களை வடிவமைப்பதற்கு வணிக ரீதியில் செயல்படும் நிறுவனங்கள் முன் வருகின்றன. அரசும் ஆதரவு தருகிறது. இத்தகைய பணிகளுக்கான மென்பொருள் வடிவமைப்பதற்குத் தற்போது கணிப்பொறிகளில் உள்ள வசதிகள் போதுமானதாகும். மேலும் சாதாரண மொழி அறிவே வடிவமைப்பாளருக்குப் போதுமானதாகும்.

ஆனால் கணிப்பொறியை மொழி ஆய்விற்குப் பயன்படுத்தும்போது மொழியில் புலமையும் செறிவும் தேவைப்படுகின்றன. இத்தகைய பணிகளைக் கணிப்பொறி வல்லுநர்களும் மொழயியலாளர்களும் இலக்கியவாதிகளும் இணைந்து செயல்பட்டு வடிவமைக்கப்பட வேண்டும்.

கணிப்பொறியில் தமிழில் உள்ளீடு செய்வதற்கும் வசதிகள் உள்ளன. அதுபோல் வெளியீட்டைப் பெறுவதற்கு உயர்ரக அச்சுப்பொறிகள் உள்ளன. ஆனால் மொழி ஆய்வுப்பணிகளைச் செய்வதற்குப் போதுமான மென்பொருள்கள் இல்லை. இதற்கான முயற்சிகள் தொடங்கப்பட்டுள்ளன.

1.7. மொழி ஆய்விற்கு தேவைப்படும் கருவிகள்

மொழியாய்விற்குத் தேவைப்படும் கருவிகளாகப் பின்வருவனவற்றைப் பட்டியலிடலாம்: அகராதிகள், சொற்களஞ்சியங்கள், சொல்வங்கி, தகவல்தளம், கிளைப்பட வங்கி, பெருந்தரவு. எந்த ஒரு மொழியாய்வுப் பணியிலும் உயர்ந்த பட்சமாகச் சொல்லடைவு வரைக் கணிப்பொறியில் தயாரிக்கலாம். சொல்லடைவில் சொற்கள் சொல்வடிவம் என்ற நிலையில் உள்ளன. இதிலிருந்து அகராதிச் சொற்களைத் தயாரிப்பதற்கு உருபனியல் பகுப்பாய்வு தேவைப்படுகிறது. இப்பகுப்பாய்வினைச் செய்வதற்குச் செயல்முறைகள் கணிதக் கூறுகளாக வடிவமைக்கப்பட வேண்டும். இது எளிய பணியல்ல. இருப்பினும் இப்பணி மொழியாய்வின் தொடக்கப் பணியாக அமைகிறது.

தமிழில் அமைந்துள்ள செய்திகளைக் கணப்பொறியில் உள்ளீடு செய்வதற்கு விசைப்பலகை பயன்படுத்தப்படுகிறது. தமிழ் இலக்கிங்களையும் பிற படைப்புகளையும் கணிப்பொறியில் தட்டச்சு செய்து உள்ளீடு செய்ய நிறைய பொருட் செலவும் கால விரையமும் ஆகும். எனவே அச்சு வடிவில் அமைந்த தமிழ்ச் செய்திகளையும் கையெழுத்து செய்திகளையும் கணிப்பொறி படித்து சொல் வடிவங்களாக மாற்ற ஒளி வருடி (Optical Character Recognizer (OCR)) என்ற தொழில் நுட்பம் பயனபடுத்தப்பட்டு ஒளி வருடிக்கான மென்பொருள்கள் உருவாக்கப்பட்டு வருகின்றன. இத்தொழில் நுட்பம் வெற்றி பெறும்பொழுது தமிழ்ச் செய்திகளை உள்ளீடு செய்வது மிக எளிய பணியாக அமையும்.

கணிப்பொறிக்குத் தரவேண்டிய கட்டளைகள் மனித மொழியில் தருவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. ஒரு மனிதன் மற்றொருவரிடம் பேசுவது போன்றும் உரையாடுவது போன்றும் கணிப்பொறியுடன் உரையாடுவதற்கு மென்பொருள் உருவாக்கப்பட்டு வருகின்றன. இப்பணிகளெல்லாம் மொழியியலாளர்களின் மற்றும் கணிப்பொறியாளர்களின் கூட்டு முயற்சியால் நடைபெறுகின்றன.

தானியங்கி மொழி பெயர்ப்பிற்கு (automatic translation) ஒரு மொழியின் இலக்கண கூறுகள் அனைத்தையும் கணிப்பொறிக்குத் தர வேண்டி உள்ளது. அதாவது கணிப்பொறிக்கு மூலமொழி மற்றும் இலக்குமொழி இரண்டிலும் தேவையான அறிவினை ஊட்டவேண்டும். இப்பணியும் கணிப்பொறி தயாரிப்பாளரின் கூட்டுமுயற்சியிலே நடைபெறுகின்றன.

1.8. மொழியாய்வில் மொழியியலாளர்களுக்குத் தேவைப்படும் திறமைகள்

மொழி ஆய்வில் மொழியியலாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- கணிப்பொறியைப் பற்றிய அடிப்படை அறிவு
- மொழியின் இலக்கணம் மற்றும் அதனை கடினப் படுத்துவதில் திறன்
- ஒரு திட்டத்தில் எப்பணிகளைக் கணிப்பொறி வழி செய்ய இயலும் என்பதற்கான திட்டம்
- ஒன்றிணைந்து பணிபுரியும் கணிப்பொறியாளரிடம் மொழி சிக்கல் பற்றியும் தீர்பு பற்றியும் எடுத்துரைக்கும் திறன்
- சிக்கலுக்கான தீர்வினைப் படிப்படியாகப் பல்வேறு நிலைகளில் செயல்படுத்துவதற்கான திறன்
- கணிப்பொறி பயன்படுத்தும் போது தோன்றும் சிக்கல்களைச் சகமொழியாளர்களிடம் காட்டி அதற்குத் தீர்வு காணும் திறன்

1.9. மொழியாய்வில் கணிப்பொறியாளர்க்குத் தேவைப்படும் திறமைகள்

மொழியாய்வில் கணிப்பொறியாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- மொழியிலும் மொழி இலக்கணத்திலும் அடிப்படை அறிவு பெற்றிருத்தல்
- மொழியாய்வுப் பணியைப் பல்வேறு சிறு சிறு பணிகளாகப் பிரித்துக் கணிப்பொறி வழி செயல்படுத்துவதற்கான மென்பொருள்களை உருவாக்கும் திறன்
- மொழியாய்வுப் பணிகளுக்கு உதவும் கணிப்பொறி வழியமைப்புகளில் தேர்ச்சி பெற்றிருத்தல்
- மொழியாளருக்குத் தேவைப்படும் மாதிரி வழியமைப்புகளை உருவாக்கி மொழியாளருக்குச் செயல் விளக்கம் தந்து அவர்களின் ஒத்துழைப்பைப் பெறுதல்
- மொழியாளர் எண்ணும் பிரச்சனைகளுக்கான தீர்வுகளை உடனுக்குடன் எதிர்பாராமல் இருத்தல். மொழியாளர்களுக்குக் கணிதப்படுத்துவதில் பயிற்சியின்மையால் உடனுக்குடன் தீர்வினைத் தர இயலாது. இந்நிலையில் அவர்களுடன் ஒத்துழைத்து தீர்வு பெற முயற்சித்தல்
- மொழியாய்வுப் பணிகளுக்கான அவ்வப்பொழுது வெளிவரும் மென்பொருள் மற்றும் கணிப்பொறி வசதிகளைத் தெரிந்துக் கொள்ளுதல்

மேற்கூறியவாறு மொழியியல் ஆய்வுப் பணிகளுக்கு மொழியாளரும் கணிப்பொறியாளரும் சேர்ந்து ஒத்துழைக்கவேண்டும். கணிப்பொறி வழி மொழியாய்வுப் பணியினை மேற்கொள்வதற்கு தேவைப்படும் மென்பொருள்களை உருவாக்க கணிப்பொறியாளர்கள், மொழியாளர்கள், இலக்கியவாதிகள் ஆகியோரின் கூட்டு முயற்சி தேவையாகும்.

1.10. தமிழில் இயற்கை மொழி ஆய்வு

1.10.1. தமிழில் இயற்கை மொழி ஆய்வுகான முன்னேற்பாடுகள்

இந்திய மொழிகள் அனைத்தும் கணிப்பொறி அறிவியலில் உரிய இடத்தைப் பெறுவதற்காக நடுவண் அரசும் மாநில அரசும் பல்கலைக் கழகங்களும் ஆராய்ச்சி நிறுவனங்களும் தனியார் நிறுவனங்களும் பல்வேறு முயற்சிகளில் ஈடுபட்டுவருகின்றன. நடுவண் அரசின் தகவல் தொழில்நுட்ப அமைச்சரகமானது இதற்கென்றே ஒரு தனித் திட்டத்தை – இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technology Development of Indian Languages – TDIL) என்ற திட்டத்தை உருவாக்கிச் செயல்பட்டுவருகிறது. இதற்கென்று நாடெங்கும் 13 பல்கலைக்கழகங்கள் மற்றும் ஆய்வு நிறுவனங்களைத் தேர்ந்தெடுத்து, பல கோடி ரூபாய் நிதி உதவி அளித்து வருகிறது. இத்திட்டத்தின் கீழ் பல இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நடைபெற்று வருகின்றன.

கடந்த ஐந்தாண்டுகளில் தமிழகத்திலும் மாநில அரசானது தமிழ்மொழியின் தொழில்நுட்ப வளர்ச்சிக்காகப் பல்வேறு திட்டங்களைச் செயல்படுத்தியுள்ளது. தமிழ் மென்பொருள் வளர்ச்சிக்கான பல்வேறு திட்டங்களை அறிவித்துச் செயல்படுத்தி வருகிறது. இதனடிப்படையில்தான் தமிழ் இணையப் பல்கலைக்கழகமும் நிறுவப்பட்டது. தமிழ் இணைய மாநாடுகள் நடைபெற்று உள்ளன. தமிழ்மென்பொருள் உருவாக்கத்திற்காக மானிய உதவியைப் பலர் பெற்றுள்ளனர்.

சென்னைப் பல்கலைக்கழக மொழியியல் ஆய்வுப் பிரிவு, அண்ணாப்பல்கலைக்கழக கணிப்பொறியியல் துறை, சென்னை குரோம்பேட்டையில் அமைந்துள்ள AU–KBC மையம், அண்ணாமலைப் பல்கலைக்கழக மொழியியல் துறை,

தஞ்சை தமிழ்ப்பல்கலைக்கழகம் மொழியியல் துறை, கோவை பாரதியார் பல்கலைக்கழக மொழியியல் துறை போன்ற பல்வேறு பல்கலைக்கழகத் துறைகளில் தமிழ் மொழியின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நிகழ்த்தப்பட்டு வருகின்றன. தமிழில் மென்பொருள் தயாரிப்பாளர்கள் இணைந்து, கணித் தமிழ்ச்சங்கம் என்ற ஒரு அமைப்பைத் தொடங்கி பல்வேறு திட்டங்களை மேற்கொண்டு வருகின்றனர்.

தமிழ்மொழி தொழில்நுட்ப வளர்ச்சியின் நிலைப்பற்றி நாம் பார்ப்பதற்கு முன்னர், உலக அளவில் மொழி தொழில்நுட்பம் (Language Technology) என்ற வளர்ச்சிக் கட்டத்தில் உள்ளது என்பதைச் சற்றுப் பார்க்கலாம். மொழி தொழில்நுட்பம் என்று அழைக்கப்படுவது என்ன? உயிரித்தொழில்நுட்பம் (Bio-technology), வேதியியல் தொழில்நுட்பம் (Chemical Technology), தோல் தொழில்நுட்பம் (Leather Technology), தகவல் தொழில்நுட்பம் (Information Technology) என்று பலத்துறைகளைப் பற்றி நம்மில் பலருக்கு அறிமுகம் இருக்கலாம். அந்த வரிசையில் தற்போது மொழித்தொழில்நுட்பம் (Language Technology) என்பதும் சேர்ந்துள்ளது. கணினித்துறையில் மனிதனின் இயற்கை மொழிகளின் பல்வேறு செயல்பாடுகளுக்கான மென்பொருள் மற்றும் வன்பொருள்களை உருவாக்கும் தொழில்நுட்பமே மொழித்தொழில்நுட்பம் என அழைக்கப்படுகிறது. எடுத்துக்காட்டாக, கணினியில் தமிழைத் தட்டச்சு செய்து, அதைச் சொல் திருத்தி, இலக்கணத் திருத்தி போன்றவற்றின் மூலம் செம்மைசெய்யும் சொல்லாளர் மென்பொருள் (Word Processor) மொழித்தொழில்நுட்பத்தின் ஒரு விளைபொருளாகும். அது போன்றே நாம் கணினியில் தட்டச்சுச் செய்வதைப் பேச்சாக மாற்றிக் கொடுக்கும் மென்பொருள் (Automatic speech Recognizer), தட்டச்சு செய்வதற்குப் பதிலாக நாம் கணினியில் பேசினால் அதை எழுத்து வடிவில் தட்டச்சு செய்து கொடுக்கும் மென்பொருள் (Text–To–Speech), இணைய தளத்தில் நமக்குத் தேவையானவற்றைத் தேடி எடுத்துக் கொடுக்கும் மென்பொருள் (Search Engine), இவையெல்லாம் மொழித் தொழில்நுட்பத்தின் விளைபொருட்களேயாகும். இவை போன்ற இயற்கை மொழிகளுக்கான மென்பொருள்

மற்றும் வன்பொருட்களை உருவாக்கும் மொழித் தொழில்நுட்பத் துறையானது பல்துறை சார்ந்த ஒன்றாகும். கணினித் துறை, தகவல் தொடர்புத்துறை, உளவியல் துறை, மொழியியல் துறை ஆகிய நான்கு துறைகளைச் சேர்ந்த வல்லுநர்களும் இணைந்து செயல்படும் ஒரு துறை இந்தத் மொழித்தொழில்நுட்பத் துறையாகும்.

1.10.2. மொழித்தொழில் நுட்பம்

மேற்கூறிய ஒவ்வொரு துறையினருக்கும் மொழித்தொழில்நுட்பத்தில் ஈடுபடுவதற்கான நோக்கங்கள் வேறுபட்டிருக்கலாம். கணிப்பொறித்துறை வல்லுநர்கள் (computer scientists) கணிப்பொறியைச் செயற்கை அறிவுத் திறனுள்ள ஒரு கருவியாக (artificial intelligent machine) மாற்றும் நோக்கத்தில் மொழித் தொழில்நுட்பத்தில் ஈடுபடலாம். ஏனென்றால் மனிதனின் அறிவுத் திறனுக்கு அடிப்படையாக அவனது இயற்கை மொழி அமைவதால், கணிப்பொறிக்கு செயற்கை அறிவுத் திறனைக் கொடுக்க முயற்சிக்கும் கணிப்பொறித் துறை வல்லுநர்கள் முதலில் கணிப்பொறிக்கு இயற்கை மொழிகளைப் புரிந்து கொள்ளும் மற்றும் பயன்படுத்தும் திறனைக் கொடுக்க முயலுகின்றனர். அதற்காக அவர்கள் இயற்கை மொழிகளின் பண்பு மற்றும் அமைப்பு என்ன, நமது மனித மூளையானது எவ்வாறு இயற்கை மொழிகளைக் கற்றுக் கொள்கிறது, பயன்படுத்துகிறது போன்றவற்றைக் கண்டறிவதற்காக மொழித்தொழில்நுட்பம் மற்றும் மொழியியல் ஆய்வுகளில் ஈடுபடுகின்றனர்.

தகவல்தொடர்புத்துறை வல்லுநர்கள் (communication engineers) நாம் ஒரு முனையில் இயற்கை மொழிகள் வழியாகப் பேசும் போது அல்லது எழுதும்போது, அதை மின்னணுச்சாதனங்கள் எவ்வாறு எவ்விதச் சொற்சேதமும் தொடர்ச்சேதமும் கருத்துச்சேதமும் இன்றிக் கொண்டுசெல்கிறது அல்லது கொண்டுசெல்லாம் என்பதை ஆய்வதற்காக மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

உளவியல் நிபுணர்களோ (psychologists) நாம் மொழி வாயிலாக மற்றவர்களோடு கலந்துரையாடும்போது என்னென்ன உத்திகளைக் கையாளுகின்றோம். இந்த உத்திகளைக் கணிப்பொறி வாயிலாக இயற்கை மொழிகளைப் பயன்படுத்தும் போது எவ்வாறு பயன்படுத்தலாம் என்பதைக் கண்டறிவதற்காகத் மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

மொழியியல் வல்லுநர்கள் (Linguists) மொழிகளின் அமைப்பு என்ன, பண்பு என்ன, எவ்வாறு பேச்சொலிகளோ எழுத்துக்களோ கருத்து அல்லது பொருளை வெளிப்படுத்துகிறது என்பவற்றையெல்லாம் கண்டறிந்து, பின்னர் இதனடிப்படையில் கணிப்பொறி வாயிலாக இயற்கை மொழிப்பயன்பாட்டை மேற்கொள்வதற்காக மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

அடிப்படையில் மொழித் தொழில்நுட்பத்தின் ஆய்வுப்பொருள் நமது இயற்கை மொழிகளேயாகும். எனவே மொழி அல்லது மொழியியல் ஆய்வின் ஒரு பிரிவே மொழித்தொழில்நுட்பமாகும். எனவே தான் மொழியியலாளர்கள் கணினிமொழியியல் (Computational Linguistics) என்ற ஒரு மொழியியல் ஆய்வுப் பிரிவை உருவாக்கி வளர்த்து வருகின்றனர். அதன் ஒரு உட்பிரிவே மொழித் தொழில்நுட்பமாகும். மொழித்தொழில்நுட்ப ஆய்வில் அல்லது வளர்ச்சியில் மொழி ஆய்வாளர்களுக்கு மிக முக்கியப் பெரும்பங்குண்டு.

உலக அளவில் கணினிமொழியியல் மற்றும் அதன் உட்பிரிவான மொழித் தொழில்நுட்ப வளர்ச்சியின் பயனாகப் பல்வேறு மொழி மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஜெர்மனியில் ஆச்சிரியப்படத்தக்க ஒரு மொழித் தொழில் நுட்ப வளர்ச்சி அண்மைக் காலத்தில் ஏற்பட்டுள்ளது. அங்கு ஆங்கிலம் மட்டுமே

தெரிந்த ஒருவர் சென்றால், அவர் தனது செல்போனில் ஆங்கிலத்திலேயே பேசலாம். மறுமுனையில் ஆங்கிலம் தெரியாத ஒரு ஜெர்மானியர் அதை ஜெர்மானிய மொழியில் கேட்கலாம். ஒரு ஜப்பானியர் தனது ஜப்பானிய மொழியில் கேட்கலாம். அதுபோன்று ஒரு ஜெர்மானியர் அல்லது ஜப்பானியர் தங்களது மொழியில் பேசினால், மறுமுனையில் ஆங்கிலம் மட்டும் தெரிந்தவர் அதை ஆங்கிலத்தில் கேட்கலாம். அதாவது உடனடிக் கணினி மொழிபெயர்ப்பு என்ற ஒரு மொழித் தொழில்நுட்ப வளர்ச்சி அங்கு ஏற்பட்டுள்ளது.

கனடா நாட்டில் வானிலையைக் கண்டறியும் கருவிகள் தரும் வானிலை விவரங்கள் ஆங்கிலத்தில் தானாக வெளிப்படுத்தும் ஒரு மொழித் தொழில்நுட்பம் உருவாக்கப்பட்டு, தற்போது செயல்படுத்தப்பட்டுவருகிறது. அதுபோன்று விமானம் மற்றும் ரயில் பயணத்திற்கான பதிவுச் சீட்டுகளைப் பதிவு செய்யப்போகின்றவர்களிடம் மனித உதவியாளர்களுக்குப் பதிலாகக் கணிப்பொறியே உரையாடி அவர்களுக்குத் தேவையான பதிவுச் சீட்டைப் பதிவு செய்துதரும் மொழித் தொழில்நுட்ப வளர்ச்சி ஐரோப்பிய நாடுகளில் ஏற்பட்டுள்ளது.

1.10.3. தமிழ்மொழித் தொழில்நுட்ப வளர்ச்சியின் இன்றைய நிலை

நமது தமிழ்மொழித் தொழில்நுட்பம் எந்த வளர்ச்சிக் கட்டத்தை எட்டடியுள்ளது என்பதை இனிப் பார்க்கலாம். 15 ஆண்டுகளுக்கு முன் தமிழ்நாட்டில் கணிப்பொறியில் தமிழைப் பார்க்கவே முடியாத நிலை இருந்தது. கணிப்பொறியில் ஆங்கிலம் போன்ற மொழிகளில் மட்டுமே தட்டச்சு செய்யவோ அல்லது பிற கணிப்பொறிப் பயன்பாடுகளை மேற்கொள்ளவோ முடிந்தது. பின்னர் தமிழ் எழுத்துருக்களை (Tamil Fonts) கணிப்பொறியில் கொண்டு வரும் முயற்சி வெற்றிபெற்றது. அதன் பயனாகக் கணிப்பொறியில் தமிழைத் தட்டச்சு செய்யும்

வளர்ச்சி ஏற்பட்டது. அதிலும் பல பிரச்சனைகள் நீடித்தன. தமிழுக்குப் பல வேறுப்பட்ட கணினித் தட்டச்சுப் பலகைகள் அறிமுகப்படுத்தப்பட்டன. மேலும் குறிப்பிட்ட ஒரு தமிழ் மென்பொருளில் தட்டச்சு செய்தலை மற்றொரு மென்பொருள் கொண்டு திருத்தவோ மாற்றவோ செய்ய இயலாத நிலை நீடித்தது. பின்னர் தமிழக அரசின் உதவியுடன் நடைப்பெற்ற தமிழ் இணையம் 99 மாநாடு ஓரளவுக்கு இப்பிரச்சனைகளைத் தீர்த்து வைத்தது. இதுபோன்ற முயற்சிகளினால் கணிப்பொறிக்கும் நமக்கும் இடையில் தமிழ்பயன்படத் தொடங்கியது. தமிழைக் கணிப்பொறியில் பார்க்கத் தொடங்கினோம். அதாவது பயன்பாட்டாளர் இடையூடகம் (User interface) என்ற அடிப்படையில் தமிழ் பயன்படத் தொடங்கியது. பல கணினி மென்பொருள்களில் தமிழ் எழுத்துக்கள் நடைபோடத் தொடங்கின. இந்த வளர்ச்சியைத் தமிழ் வாயிலாகக் கணிப்பொறியின் பயன்பாடு என்று அழைக்கலாம்.

ஆனால் ஆங்கில மொழிக்கு உருவாக்கப்பட்டுள்ள சொல்லாளர் (Word Processor), பேச்சறிவான் (Speech Recognizer), எழுத்து–பேச்சு மாற்றி (Text–To–Speech transferor), இணையத்தேடுவான் (Speech Engine) போன்ற பலமொழிசார் மென்பொருள்கள் தமிழுக்கு இன்று இல்லை. ஆங்கிலத்தைக் கணிப்பொறியில் தட்டச்சு செய்யும் போது, சொற்பிழை அல்லது இலக்கணப் பிழை ஏற்பட்டால் தானாக அவற்றைத் திருத்தித்தரும் சொல்லாளர்கள் பல உள்ளன. ஆனால் இன்னும் தமிழுக்கு ஒரு சரியான சொல்திருத்தியையும் இலக்கணத் திருத்தியையும் கொண்ட முழுமையான சொல்லாளர் உருவாக்கப்படவில்லை. இதுபோன்ற தமிழ் மென்பொருள்களைத் தமிழுக்கான மென்பொருட்கள் (Software For Tamil) என்று அழைக்கலாம். இவ்வகை மென்பொருட்கள் உருவாக்கப்படுவதில் என்ன தடைகள் என்பதைப் பின்னர் பார்க்கலாம்.

மருத்துவத்துறை மென்பொருட்கள் உருவாக்கப்படவேண்டுமென்றால், முதலில் குறிப்பிட்ட மருத்துவத்துறையில் அறிவு உடையோர் அத்துறையில் ஆராய்ச்சி மேற்கொண்டு, மருத்துவ உண்மைகளைத் தெளிவுப்படுத்தும்போதுதான், அவற்றை அடிப்படையாகக் கொண்டு, மருத்துவர்களின் உதவியோடு கணிப்பொறித்துறை வல்லுநர்கள் மருத்துவ மென்பொருட்களை உருவாக்கமுடியும். அவ்வாறு இல்லாமல் கணிப்பொறித்துறை வல்லுநர்களுக்கும் தான் உடம்பு, உயிர் இருக்கிறதே, அவர்களுக்கு அவைபற்றித் தெரியாதா, அவர்களே அதை அடிப்படையாகக் கொண்டு மருத்துவ மென்பொருட்களை உருவாக்கக்கூடாத என்று யாரும் கேட்பார்களா? கேட்க மாட்டார்கள். அதுபோல ஒரு மொழிக்கான மென்பொருளை உருவாக்கவேண்டும் என்றால், மொழியறிஞர்கள் அக்குறிப்பிட்ட மொழியின் அமைப்பு, பண்பு, செயல்பாடு போன்றவற்றை ஆராய்ந்து வெளிப்படுத்தும்போது தான், அவற்றை அடிப்படையாகக் கொண்டு கணிப்பொறித்துறை வல்லுநர்கள் அம்மொழிக்கு மென்பொருட்களை உருவாக்க இயலும். அவ்வாறு இல்லாமல் அம்மொழிபேசும் கணிப்பொறித்துறை வல்லுநர்கள் தாங்களே அம்மென்பொருட்களை உருவாக்க வேண்டியதுதானே என்று கேட்கக் கூடாது. அவ்வாறு கேட்பது தவறு. மொழி வல்லுனர்களின் துணையின்றி கணிப்பொறி வல்லுனர்கள் மொழிசார் மென்பொருள்களைச் செம்மையாக உருவாக்கவியலாது என்பது தான் உண்மை.

தமிழுக்கு ஒரு நல்ல சொல்திருத்தி (Spell Checker) உருவாக்கப்படவேண்டும் என்றால், தமிழ்ச் சொற்களின் பண்பு மற்றும் அமைப்பைப் பற்றி உண்மைகளைத் தெளிவாகத் தமிழ்மொழி அறிஞர்கள் முன்வைக்கும்போதுதான், அதை அடிப்படையாகக் கொண்டு, கணிப்பொறித்துறை வல்லுநர்கள் சொல்திருத்தி மென்பொருளைத் தமிழுக்கு உருவாக்க முடியும். தமிழில் பெயர்ச்சொற்கள் எந்தெந்த விகுதிகளை எடுக்கும், எந்த வரிசையில் எடுக்கும், எவ்வாறு அவற்றை இணைத்து ஒரு முழுச் சொல்லை உருவாக்குவது, எவ்வாறு தமிழ்ச்சொற்களைப் பகுபத உறுப்புகளாகப்

பிரிப்பது போன்ற உண்மைகளை எல்லாம் தெள்ளத்தெளிவாக, அறிவியல் பூர்வமாகத் தமிழ் மொழி அறிஞர்கள் எடுத்துச்சொன்னால், கணிப்பொறித்துறை வல்லுநர்கள் அவற்றை அடிப்படையாகக் கொண்டு சொல்திருத்தி மென்பொருட்களை உருவாக்க இயலும்.

இந்த இடத்தில் ஓர் உண்மையைத் தெளிவுப்படுத்த வேண்டியது அவசியம். மனித மூளை மொழியைப் புரிந்துகொள்ளும் முறை வேறு, கணிப்பொறி புரிந்துகொள்ளும்முறை வேறு. அதற்குக் காரணம், மனித மூளை ஒரு உயிர் வேதியியல் பொருள்; ஆனால் கணிப்பொறியில் உள்ள சிப் (chip) ஒரு மின்னணுப்பொருள். எனவே எந்த ஒரு அறிவையும் மனித மூளைக்குக் கொடுக்கும் முறை வேறு, மின்னணுச் சிப்புக்குக் கொடுக்கும் முறை வேறு, இவ்வேறுபாட்டை மனதில் கொண்டு மொழி அறிஞர்கள் தாங்கள் கண்டறிந்த உண்மைகளைக் கணிப்பொறிக்குப் புரியும் வகையில் மாற்றி அமைத்துகொடுத்தால் தான் கணிப்பொறி வல்லுநர்களின் வேலை எளிதாகும். இதுபோன்ற இலக்கணங்களைக் கணினி இலக்கணம் (Computational Grammar) என்று அழைப்பார்கள்.

ஆங்கிலம் போன்ற மொழிகளுக்குப் பல்வேறு நோக்கில் பல்வேறு கணினி இலக்கணங்கள் உருவாக்கப்பட்டுள்ளன. எனவேதான் அவற்றிற்கு நல்ல சொல்லாளர்கள், பேச்சறிவான்கள், எழுத்து–பேச்சு மாற்றிகள், இணையத் தேடுவான்கள், இயந்திர மொழிபெயர்ப்பிகள் போன்ற மென்பொருட்கள் நிலவுகின்றன.

இவ்வாறு கணிப்பொறி நோக்கில் மொழியாய்வு மேற்கொள்ளப்படுவதுதான் கணினி மொழியியல் என்று அழைக்கப்படுகிறது. அதன் உட்பிரிவே மொழித் தொழில்நுட்பம் ஆகும். எனவே தமிழ்க் கணினி மொழியியல் வளர்ச்சி அடையும்போதுதான் தமிழுக்கான பல மென்பொருட்கள் உருவாகும். இதுவே

கணிப்பொறியில் உயர்நிலைத் தமிழ்மொழிப் பயன்பாடாகும். தற்போது தமிழ் எழுத்துருக்களை (Fonts) உருவாக்கி, கணிப்பொறியில் பிற மென்பொருட்களோடு உறவாடும் நிலைதான் (User Interface) ஏற்பட்டுள்ளது. அடுத்தக் கட்ட நிலைக்கு நாம் செல்ல வேண்டும். இந்தி, தெலுங்கு, வங்காளம், மராத்தி போன்ற மொழிகளுக்கு இதுபோன்ற உயர்நிலை மென்பொருட்கள் உருவாக்கப்பட்டுள்ளன.

கணினிமொழியியலும் அதன் உட்பிரிவான மொழித் தொழில்நுட்பமும் இன்று உலகளவில் நன்கு வளர்ச்சி அடைந்துள்ளன. இதற்கென்று தனித்துறைகள் பல்வேறு பல்கலைக்கழகங்களில் உருவாக்கப்பட்டுச் செயல்பட்டுவருகின்றன. தமிழ்மொழியிலும் கணினிமொழியியல் கோட்பாடுகளின் அடிப்படையில் பல ஆய்வுகள் மேற்கொள்ளப்படவேண்டும். தமிழுக்குச் சொற்பகுப்பான் (Morphological Parser), சொல் உருவாக்கி (Word Generator), தொடர்பகுப்பான் (Syntactic Parser), தொடர் உருவாக்கி (Sentence Generator) போன்ற அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படுவதும் பௌதிக ஒலியியல் அடிப்படையிலான தமிழ்ப்பேச்சொலி ஆய்வு, பொருண்மையியல் (Semantics) அடிப்படையிலான ஆய்வு, விரிதரவு மொழியியல் ஆய்வு (Corpus Linguistics) போன்றவை மேற்கொள்ளப்படுவதும் கணினித் தமிழ் வளர்ச்சிக்கு இன்றியமையாதன. இவை இல்லாமல் தமிழுக்கு உயர்நிலையில் மென்பொருட்கள் உருவாக்க இயலாது.

கணிப்பொறிமயமாக உள்ள இன்றைய உலகில், உயர்நிலைத் தமிழ்மென்பொருட்கள் தமிழுக்கு உருவாக்கப்படாமல் உலக அளவில் மற்ற மொழிகளுக்கு இணையாகத் தமிழ் அடுத்த கட்ட வளர்ச்சியை எட்டவியலாது. அச்சு இயந்திரம் தோன்றியபோது, தமிழ் அதில் இடம் பெற்ற காரணத்தால்தான் இன்று தமிழ் எழுத்துலகில் நீடித்து வளர்ந்து வந்துள்ளது என்பதை நாம் மறந்துவிடலாகாது. அதுபோன்று கணிப்பொறியில் உரிய இடத்தைத் தமிழ் பெற்றால்தான் அடுத்த கட்ட

வளர்ச்சிக்குத் தமிழ் செல்லும் என்பதை நாம் நினைவில்கொள்ள வேண்டும். தமிழுக்கு எழுத்துருவை உருவாக்கிக் கணிப்பொறியில் பார்ப்பதோடு நிறுத்திக்கொள்ளாமல், பல்வேறு உயர்நிலைத் தமிழ்மென்பொருள் உருவாக்கத்தில் நாம் கவனம் செலுத்தவேண்டும்.

எனவே, தமிழ் ஆசிரியர்களும் ஆய்வாளர்களும் மாணவர்களும் கணினிமொழியியல் மற்றும் மொழித் தொழில்நுட்பம் அடிப்படையில் தமிழ் ஆய்வை மேற்கொள்ளவேண்டும். உயர்நிலைத் தமிழ் மென்பொருட்கள் உருவாக்கத்திற்குத் தேவையான தமிழுக்கான அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படவேண்டும்.

இயல் 2

கணிப்பொறி மூலம் தமிழ்மொழி ஆய்வு

2.1. மொழி ஆய்வும் கணிப்பொறியும்

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப்பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களை பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. வெறும் கணிக்கும் பொருளாக இருந்த கணிப்பொறி தரவுகள் சேமித்து வைக்கும் கருவியாக, தகவல் தொடர்பு கருவியாக தரவுகளை தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகிறது. மொழி செயல்பாட்டிற்கும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவில் உள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கானச் சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத்தட்டு, பதிவுநாடா போன்றவற்றில்) சேகரித்து வைத்துத் தேவையானபோது பயன்படுத்திக் கொள்ள இயலும்.

மொழிக்கான தரவுகள் பலவகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவோ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். இவைகள் உள்ளடங்கிய ஒரு பனுவலாகவும் (text) அமையலாம். வெவ்வேறு வகைப்பட்ட உரைகளிலிருந்து தெரிந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெரும்தரவு (corpus) என்று அழைக்கப்பெறும். தழிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக இன்றியமையாதப் பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய உரைகள், பல நடையியல் கூறுகளை உள்ளடக்கிய உரைகள், வட்டார வழக்கு, கிளை மொழி வழக்கு, சமுதாயமொழி வழக்கு, பேச்சு வழக்கு, இலக்கியவழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய உரைகள் எல்லாம் ஒருசேர பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம். 1. எழுத்து வடிவில் அமையும் பெருந்தரவு (Written corpus) மற்றொன்று பேச்சு வடிவில் அமையும் பெருந்தரவு (speech corpus). இத்தகைய பெருந்தரவுகள் மொழியாய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவிப் புரிகின்றன.

தரவுகள் கணிப்பொறியால் ஆயப்பட்டு தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன்தட்டு, குறுந்தட்டு, பதிவுநாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும்போது பயன்படுத்திக் கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையத்தளத்தில்

பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல்மொழியைப் பயன்படுத்தி தேவையான தகவல்களைப் பெறுதலாகும். இதற்கு தேவை இயல்மொழி இடைமுகப்பு (Natural language Interface) இதனால் இயல்மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டு தேவையான தகவல்களைப் பெற இயலும். தமிழ்மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ்மொழியைப் பயன்படுத்தி தகவல்களைப் பெற உதவும். இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல்தேடும் பொறி (search engine) தகவல்களைப் பெற உதவும் முக்கியக் கருவியாகும். தமிழுக்கு இத்தகைய தகவல்தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு பெருமளவில் வெற்றியும் அடைந்துள்ளது.

கேள்வி-பதில் அமைப்புகள் கணிப்பொறியுடன் உரையாடும் அமைப்புகள் போன்றவை இயல்மொழி இடைமுகப்பை அடிப்படையாகக் கொண்டு அமைக்கப்பட்டிருந்தால் பாமரரும் பயன்படுத்தி தகவல்களைப் பெற இயலும். எடுத்துக்காட்டாக, தமிழ் இலக்கிய நூல் ஒன்றை விடுவிரலாக தந்து இயல்மொழி முகப்பு மூலம் தேவையான பதில்களை அல்லது தகவல்களை பெறும்படியாக கணிப்பொறியின் பயன்பாடு அமையலாம். விவசாயிகளுக்கான தரவு மையங்களிலிருந்து இயல்மொழியின் வழி வேண்டிய தகவல்களைப் பெறுதலும் இத்தகைய பயன்பாடாகும்.

மொழி உரைகளைப் படிநிலை (ஒலியனியலி நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழி அலகுகளாக பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துக் கொள்ள கணிப்பொறியைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக் கொள்வோம்:

பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்

பள்ளி மாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும், *பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத்தொடர் என்றும், மேலும் *பள்ளி மாணவன், மரம், பலாப்பழம், ரகுமான்* என்பன பெயர்கள் என்றும், *பறி* என்பது வினை என்றும், *பெரிய* என்பது பெயரடை என்றும், *கவனமாக* என்பது வினையடை என்றும், *இலிருந்து* பின்னுருபு என்றும், *ஐ* என்பது வேற்றுமையுருபு என்றும், -*த்*- என்பது இறந்தகால இடைநிலை என்றும், *ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிபொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களை பகுத்தாய்வுச் செய்து தந்துவிடும்.

ஒரு காலகட்டத்தில் மொழியலார் கள ஆய்வு செய்து தகவலாளிகளின் உதவியால் ஆயப்படாத மொழிகளின் தரவுகளைச் சேகரித்து அதன் மீது சில மொழியியல் கொள்கைகளையும் கோட்பாடுகளையும் பயன்படுத்தி சில கருதுகோள்களின் அடிப்படையிலும் உற்றுநோக்கல் மூலமும் அம்மொழிகளின் ஒலியனியல், உருபனியல், தொடரியல் அமைப்புகளைக் கண்டபிடிக்க முயலுவர். இது பண்புசார் மொழியியல் (qualitative linguistics) எனப்படும். இப்போது கணிப்பொறியின் வரவால் அத்தகைய நிலை மாறி மொழிகளின் விரிதரவுகள் உருவாக்கப்பட்டு அதன்மீது புள்ளியில் நெறிமுறைகளின் அடிப்படையில் மேற்சொன்ன மொழி அமைப்புகள் கண்டுபிடிக்கப்படும். இதை அளவுசார் மொழியியல் (quantitative linguistics) என்பர். தற்போது மொழி ஆய்வுகள் எல்லாம் கணினியைப் பயன்படுத்தி விரிதரவுகள் மீது செய்யப்படும் ஆய்வுகளாக மாறியுள்ளன.

மொழி உரையை இடு வரலாகத் (input) தந்து தக்க வழியமைப்புகளும் (programs) கட்டளைகளும் தந்தால் மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (out put) பெறலாம். கணிப்பொறியைக் கொண்டு மொழி உரைகளை வாக்கியங்களாகவும் வாக்கியங்களைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக்கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காக தொடரியல் ஆய்வி (Syntactic analyzer), உருபனியல் ஆய்வி (Morphological anayazer), ஒலியனியல் ஆய்வி (Phonological analyzer), பொருண்மையியல் ஆய்வி (Semantic anayazer) என்ற ஆய்விகளைத் தயாரிக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் தயாரிக்கப்பட்டுப் பயன்பாட்டில் உள்ளது. இவ்வாறு மொழிக் கூறுகளைப் கணிப்பொறி வழி பகுப்பாய்வது குறித்தும் தமிழில் இதற்கான முயற்சிகள் குறித்தும் இங்கு விளக்கப்படும். மேலும் இவ்வாய்விகள் அடிப்படையிலான மொழிசார் மென்பொருள்கள் அல்லது மொழிக் கருவிகள் உருவாக்குவது குறித்தும் இங்கு விளக்கப்படும்.

2.2. உரையிலிருந்து பேச்சு

பேச்சு சொல் பேச்சின் சிறு அலகுகளால் ஒருங்கிணைக்கப்பட்டது என்ற ஒலி அடிப்படையிலான எழுதும் ஒழுங்குமுறையில் உள்ளுறையும் கருத்து ஒலியனியலின் எல்லாத் தற்காலக் கோட்பாடுகளையும் உள்ளடக்கும் Ur-கோட்பாடு ஆகும். நாம் இங்கு தற்காலப் பேச்சறிதல் மற்றும் பேச்சுருவாக்கத் தொழில்நுட்பத்தையும் கணினி ஒலியனியல் (Computational Phonolgy) என்ற தொடர்புள்ள மொழியியலின் கிளைளையும் அறிந்துகொள்ள அடிப்படை உள்ளறிவையும் வழிமுறை வரைவுகளையும் அறிமுகப்படுத்த வேண்டும்.

தானியங்கு பேச்சறியும் முக்கியமான செயல்பாடு ஒலியியக்க அலைவடிவை உள்ளீடாக ஏற்றுச் சொற்களின் கோர்வையை வெளியீடாக உற்பத்தி செய்யும். எதிரிடையாகப் உரையிலிலிருந்து பேச்சுருவாக்கம் என்ற முக்கியச் செயல்பாடு உரை சொற்களின் கோர்வையை உள்ளீடாக ஏற்று ஒலியியக்க அலைகளை வெளியீடாகப் பெறுவதாகும். பேச்சைப் புரிந்து கொள்ளுதல் மற்றும் உருவாக்குதல் இவற்றின் பயன்பாடு பல்வகைப்பட்டதாகும்: தானியங்கு கேட்டெழுத்துதல் (dictation)/ ஒலிபெயர்ப்பு, கணிப்பொறிக்குப் பேச்சு அடிப்படையிலான இடைமுகங்கள் மற்றும் தொலை பேசிகள், பேச்சுக் குறையுள்ளவர்களுக்குக் குரல்சார்ந்த உள்ளீடு மற்றும் வெளியீடு என்பனவாகும்.

இங்கு நாம் பேச்சறிவான் மற்றம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள் என்பனவற்றில் கவனம் செலுத்துவோம்: எவ்வாறு சொற்கள் ஒலிகள் என்ற தனிப்பட்ட பேச்சு அலகின் அடிப்படையில் உச்சரிக்கப்படுகிறது? பேச்சறியும் ஒழுங்குமுறை அது புரிந்துக் கொள்ள இயலும் ஒவ்வொரு சொல்லுக்கும் உச்சரிப்பை வேண்டும். உரையிலிருந்து பேச்சு ஒழுங்குமுறை ஒவ்வொரு சொல்லுக்கும் ஒரு உச்சரிப்பை வேண்டும். நாம் இங்கு உச்சரிப்பை விளக்க ஒலி நெடுங்கணக்குகளை (phonetic alphabets) ஒலியில் களத்தின் பகுதியாக அறிமுகப்படுத்த வேண்டி வரும். மேலும் நாம் ஒலிப்புமுறை ஒலியியல் (articulatory phonetics) என்பதையும் அறிய வேண்டி வரும். இது எவ்வாறு வாயிலுள்ள ஒலிப்பான்களால் பேச்சொலிகள் உருவாக்கப்படுகிறது என்பதை ஆயும்.

ஒரு குறிப்பிட்ட ஒலி எப்பொழுதும் ஒவ்வொரு சூழலிலும் ஒரேமாதிரி உச்சரிக்கப்படுகிறது என்றிருக்குமானால் உச்சரிப்பை மாதிரிப்படுத்துவது எளிதாகும். ஆனால் இது அவ்வாறில்லை. எடுத்துக்காட்டாக, [த்] என்ற ஒலி வேறுப்பட்ட ஒலிச்

சூழல்களில் மிக வேறுபட்டு மெய்ப்படுத்தம் செய்யப்படுகிறது. மொழியியலில் ஒலியனியல் (phonology) களம் வேறுபட்ட சூழலில் ஒலிகள் வேறுபட்டு மெய்படுத்தம் செய்யப்படுகின்றது என்பதையும் எவ்வாறு ஒலிகளின் இவ்வொழுங்குமுறை மீதி இலக்கணத்துடன் தொடர்புடையது என்பதையும் விளக்குகிறது. நாம் இவ்வேறுப்பட்ட மெய்ப்படுத்தங்களை விளக்க ஒலியனியல் விதிகளை எழுத வேண்டும்.

கணினி வழி ஒலியனியல் ஆய்வு கணினி ஒலியனியல் (computational phonology) என்று அறியப்படும். கணினி ஒலியனியலின் முக்கியமான பகுதி ஒலியனியல் விதிகளை மாதிரிப்படுத்தும் கணினி இயங்குமுறைகளின் ஆய்வாகும். நாம் எவ்வாறு எழுத்துக்கூட்டல் மாற்றிகள் (spelling-rule transducers) ஒலியனியலை மாதிரிப்படுத்த இயலும் என்று காட்ட வேண்டும். பின்னர் ஒலியனியல் கற்றலின் (phonological learning) கணினி மாதிரிகளை விளக்க வேண்டும்: எவ்வாறு ஒலியனியல் விதிகள் இயந்திரக் கற்றல் வழிமுறைவரைவுகளால் தானாக ஊக்கப்படுத்த இயலும். இறுதியாக நாம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளில் ஒரு முக்கியமான சிக்கலுக்கு வேண்டி ஒலியனியலின் மாற்றி அடிப்படையிலான மாதிரியைப் (transducer-based model) பிரயோகிக்க வேண்டும்: எழுத்துக்களின் கோர்வைகளிலிருந்து ஒலிகளின் கோர்வைகளைப் பொருத்துதல். நாம் முதலில் பெரிய ஒலிப்பு அகராதியை உருவாக்குவதில் உள்ள சிக்கல்களை அளவிட வேண்டும். பின்னர் எவ்வாறு மாற்றி அடிப்படையிலான சொற்களஞ்சியங்களும் எழுத்துக்கூட்டல் அலகுகளும் எழுத்திலிருந்து உச்சரிப்புக்குப் பொருத்துவதற்கு உச்சரிப்புகளால் பெரிதுபடுத்தப்பட வேண்டும் என்று காட்ட வேண்டும்.

2.2.1. பேச்சு ஒலிகளும் ஒலி பெயர்ப்பும்

சொற்களின் உச்சரிப்பை ஆய்வது ஒலியியலின் ஒரு பகுதியாகும். ஒலியனியல் உலக மொழிகளில் பயன்படுத்தப்படும் பேச்சு ஒலிகளின் ஆய்வாகும். நாம் ஒரு சொல்லின் உச்சரிப்பை ஒலிகள் அல்லது கூறுகள் இவற்றை உருப்படுத்தம் செய்யும் குறியீடுகளின் கோர்வையாகும் என்று மாதிரிப்படுத்தலாம். நாம் ஒலிகளை ஆங்கிலம் போன்ற நெடுங்கணக்கு மொழியிலுள்ள எழுத்துக்களுடன் ஒற்றுமையுள்ள ஒலிக் குறியீடுகளால் உருப்படுத்தம் செய்வோம். எடுத்துக்காட்டாக, [ல்] என்பதால் உருப்படுத்தப்படும் ஒலி ல் என்ற எழுத்துடன் பொருந்தும். அதேபோல ப் என்பதால் உருப்படுத்தப்படும் ஒலி [ப்] என்ற ஒலியுடன் பொருந்தும். ஒலிகளுக்கு எழுத்துக்களைக் காட்டிலும் கூடுதல் வேறுபாடு இருக்கும். மீக்கூறு (prosody) இசைமை (pitch) மற்றும் மாத்திரை (duration) என்பனவற்றை உள்ளடக்கும்.

2.2.2. ஒலியன்களும் ஒலியன் விதிகளும்

ஒலிகள் வேறுபட்ட சூழல்களில் வேறுபட்டு உருவாக்கப்படும். ஒலியன்கள் தமக்குள் சூழல் அடிப்படையில் மாற்றொலிகளின் ஒரு குழுமத்தால் உருப்படுத்தம் செய்யப்படும். ஒலியனுக்கும் அதன் மாற்றொலியன்களுக்கும் இடையில் உள்ள உறவு ஒலியன்களின் விதியால் பெறப்படும். எடுத்துக்காட்டாக, தமிழில் \க்\ என்ற ஒலியனை உருப்படுத்தம் செய்யப்படும் மூன்று மாற்றொலியன்களைப் பின்வருமாறு ஒலியனியல் விதியால் விளக்கலாம்:

k ⟶ g/N
k ⟶ x/V-V
k ⟶ k/ பிற சூழல்களில்

அதாவது மூக்கொலி முன் கடையண்ணக் குரலிலா ஒலியானது கடையண்ணக் குரல் ஒலியாகவும் இரு உயிர்களுக்கிடையில் கடையண்ண உரசொலியாகவும் பிற சூழல்களில் குரலிலா அடைப்பொலியாகவும் உருப்படுத்தம் பெறும்.

2.2.3. ஒலியனியல் விதிகளும் முற்றுநிலை மாற்றிகளும்

ஒலியனியல் விதிகளை முற்றுநிலை மாற்றிகளால் நடைமுறை படுத்தலாம். ஒலியனியல் விதிகளை மெய்ப்படுத்தம் செய்ய முற்றுநிலைத் தானியங்கியை பலவழிகளில் பயன்படுத்தும் கணினி ஒலியனியலின் வேறுப்பட்ட மாதிரிகள் இருக்கின்றன. இவற்றில் ஒன்று நாம் கணினி உருபனியலில் பார்க்க இருக்கும் கொஸ்கென்னியின் இருமட்ட உருபனியலாகும் (two-level morphology of Koskenniemi(1983)). இருமட்ட உருபனியல் சொல் வடிவுக்கும் புற வடிவுக்கும் இடையே உள்ள பொருத்தத்தின் மீது முற்றுநிலை நல்லுருவாக்கக் கட்டுப்பாடுகளாக மாதிரி செய்யும் உருபனியல்/ஒலியனியல் கோட்பாடாகும்.

கணினி ஒலியனியலின் உயர்வான சிக்கல்களை நேரிட ஒலியனியலின் இருமட்ட அல்லது முற்றுநிலை மாதிரியின் கூடுதல் சிறப்பான உபாயம் பயன்படுத்தப்படும். ஒலியனியல் விதிகளைப் பொறியால் கற்கச் செய்யலாம். ஒரு பரப்பிலிருந்து சிறிது தரவும் பிற தகவல்களும் தரப்படுகையில் அது அப்பரப்பிற்குத் தானாக ஒரு மாதிரியைத் தூண்டுவது பொறி கற்றல் ஒழுங்குமுறையின் வேலையாகும். இவ்வாறு ஒலியனியல் விதிகளைக் கற்கவேண்டி ஒரு ஒழுங்குமுறைக்கு அதிலிருந்து தூண்ட சொற்களின் (புற வடிவங்களின்) குறைந்தது ஒரு குழும விதிகள் தரப்படும். ஒரு கண்காணிக்கப்பட்ட வழிமுறைவரைவு (supervised algorithm) இத்தரவின் சிறுபகுதிக்கு விடைகள் தரப்படுகையில் இவ்விடைகளைப் பயன்படுத்தி அது முன்னர் பார்க்காத புதிய தரவைப் பொதுமையாக்கம் செய்யவியலும் ஒரு மாதிரியைத் தூண்டும். கண்காணிக்கப்படாத வழியமைப்பு (unsupervised algorithm) இதைத் தரவிலிருந்தே செய்கின்றது. கண்காணிக்கப்படாத வழியமைப்புகள் வகுப்புகளுக்குச் சரியான புலக்குறிகளைப் பார்க்காமல் இருந்தாலும், அவைகள் உருவாக்க வேண்டிய விதிகளின் அல்லது மாதிரிகளின் இயல்பைப் பற்றி குறிப்பைத் தர இயலும். எடுத்துக்காட்டாக,

தானியங்கி வடிவில் மாதிரிகள் இருக்கும் என்ற அறிவு ஒருவகைக் குறிப்பாகும். இம்மாதிரியான குறிப்புகள் கற்றல்சார் தவறானகருத்துகள் (learing bias) எனப்படும்.

2.2.4. உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல்

நாம் ஒரு மொழியின் ஒலிகளின் பட்டியலைத் தெரிந்துக் கொள்ளவும் எவ்வாறு ஒலியனியல் விதிகளை மாதிரிப் படுத்துவது என்பதை அறிந்துகொள்ளவும் செய்தால் எழுத்திலிருந்து அல்லது உரைச் சொல்லிருந்து அதன் உச்சரிப்புக்குப் பொருத்தும் சிக்கலை ஆயத் தயார் என்பதாகும்.

2.2.4.1. உச்சரிப்பு அகராதிகள்

உச்சரிப்பு அகராதிகள் உரையிலிருந்து பேச்சுக்குப் பொருத்த மிக பயனுள்ள கருவியாகும். எளிய உச்சரிப்பு அகராதிகள் சொற்களின் பட்டியலையும் அவற்றின் உச்சரிப்பையும் தரும். எ.கா.

Word	Pronunciation	Word	Pronunciation
cat	[kæt]	goose	[gus]
cats	[kæts]	geese	[gis]
pig	[pɪg]	hedgehog	[‘hɛʤ.hɔg]
pigs	[pɪgz]	hedgehogs	[‘hɛʤ.hɔgz]
Fox	[fax]		
Foxes	[‘fak.sɪz]		

PRONLEX, CMUdict, and CELEX என்ற இணையத்தள உச்சரிப்பு அகராதிகள் பொதுவாகப் பயன்படுத்தப்படுகின்றன. இவைப் பேச்சைத் தெரிந்து கொள்கைக்குப்

பயன்படுத்தப்படுகிறது. இவற்றைப் பேச்சு கூட்டிணைப்புக்கும் பயன்படுத்த வேண்டி சுவீகரித்துக் கொள்ளலாம்.

2.2.4.2. அகராதியை நோக்குவதற்கு அப்பாற்பட்ட உரை ஆய்வு

ஆரம்பகால உரைகளிலிருந்து பேச்சாய்வுகள் உரையிலிருந்து ஒலிகளுக்குப் பொருத்த உச்சரிப்பு அகராதிகளை சார்ந்திருந்தன. முன்னர் கூறியபடி உரைகளிலிருந்து ஒலிகளாகப் பொருத்தமாற்றம் செய்யும் ஒரு வழி ஒவ்வொரு சொற்களின் உச்சரிப்பையும் உச்சரிப்பு அகராதியைப் பார்த்தறிந்து ஒலிகளின் கோர்வையை அகராதியிலிருந்து படிப்பதாகும். இந்த நெறிமுறை ஏற்கெனவே அகராதியில் தரப்பட்ட சொற்களுக்குப் பொருத்தும். ஆனால் பலச்சொற்கள் முன்னரே அகராதியில் தரப்படாததற்கான சூழல் சாத்தியம் இருக்கின்றன. இத்தகைய நிலையில் நாம் அகராதியில் தரப்படாத சொற்களின் உச்சரிப்பைத் தெரிந்துக்கொள்ள நிகழ்வுசாத்திய நுட்பங்களைப் (probabilistic techniques) பயன்படுத்த வேண்டிவரும்.

2.2.1.4.3. முற்றுநிலைமாற்றி அடிப்படையில் உச்சரிப்பு அகராதி

உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளின் தொடக்கக்கால உச்சரிப்பு மாதிரி எழுத்திலிருந்து ஒவ்வொரு விதியும் எவ்வாறு ஒரு எழுத்து அல்லது எழுத்துக்களின் இணைப்பு ஒலிகளாக பொருத்தமாற்றம் செய்யப்பட்டன என்று குறிப்பிட்டுக் காட்டியது. கீழே விற்றன் (Witten, 1982) என்பவரின் விதி அடிப்படை வடிவு தரப்பட்டுள்ளது (Jurafsky and Martin, 2000: 125):

Fragment	Pronunciation
- p -	[p]
- ph -	[f]

- phe - [fi]
- phes - [fiz]
- place - [pleɪs]
- placi - [pleɪsi]
- plement - [plɪmɛnt]

ஸ்.ப்ரோட் (Sproat) மற்றும் அவருடன் சேர்ந்த பிறரும் உரையிலிருந்து பேச்சுக்கு மாற்றிகளைப் பயன்படுத்தினார்கள். நாம் அவர்களின் மாற்றி அணுகுமுறையை ஐந்து பிரிவுகளாகப் பிரிக்கலாம்:

1. அகராதியிலுள்ள தனிச் சொற்களுக்கும் உருபன்களுக்கும் உச்சரிப்பை உருப்படுத்தம் செய்யும் முற்றுநிலை மாற்றி
2. உருபன்களின் சாத்தியமானத் தொடர்ச்சியை உருப்படுத்தம் செய்யும் முற்றுநிலைத் தானியங்கிகள்
3. ஒவ்வொரு உச்சரிப்பு விதிகளுக்கும் தனியான முற்றுநிலை மாற்றி
4. பெயர்களின் மற்றும் சுருக்கப்பெயர்களின் உச்சரிப்பை மாதிரிப்படுத்தப் பயன்படுத்தப்படும் உள்ளுணர்வு மற்றும் எழுத்திலிருந்து-ஒலி விதிகள்/ மாற்றிகள்
5. எந்த பிற அறியாத சொற்களுக்கும் வழுநிலை எழுத்திலிருந்து-ஒலி விதிகள்/மாற்றிகள்

உரையிலிருந்துப் பேச்சுக்கான உரையிலிருந்து ஒலிக்குப் பொருத்தமாற்றம் (Jurafsky and Martin, 2000:127):

Orthographic Lexicon	Lexicon
Regular Nouns	
cat	c/k a|æ t|t
fox	f|f o|ɑ x/ks
dog	d|d o|ɑ g|g
Irregular Singular Nouns	
goose	g|g oo|u s|s e|ɛ
Irregular Plural Nouns	
go:e o:e se	g| g oo|u: ee|i s|s e|ɛ

மேற்சொன்னப் படத்தில் அகராதிக்கு இரண்டு நிலைகள் இருக்கின்றன. ஒன்று உள்ளுறையும் சொல்சார் மட்டம்; மற்றொன்று இடைப்பட்ட மட்டம். இதில் இடைப்பட்ட மட்டத்தை புற மட்டத்திற்கு பொருத்தமாற்றம் செய்ய எழுத்துக்கூட்டல் விதிகளையும் உச்சரிப்பு விதிகளையும் பயன்படுத்தும் மாற்றிகளைச் சேர்க்க வேண்டும். இது வேறுபட்ட எழுத்துகூட்டல் விதிகளையும் ஒலியன் விதிகளையும் உள்ளடக்கும்.

சொற்களஞ்சியம் மற்றும் ஒலியனியல் விதிகள் மற்றும் வரிவடிவ விதிகள் என்பனவற்றைச் சொல் உருப்படுத்தத்திற்கும் புற உருப்படுத்தத்திற்கும் இடையில் பொருத்த மாற்றம் செய்யப் பயன்படுத்தலாம். சொற்களஞ்சிய முற்று நிலை மாற்றிகள், அடிகள் மற்றும் ஒலியனியல் பண்புக் கூறுகளைக் கொண்ட சொல் மட்டத்திற்கும் உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் இடைப்பட்ட மட்டத்திற்கும் இடையில் பொருத்தமாற்றம் செய்கிறது. பின்னர் ஒரு தனி எழுத்து விதிக் கட்டுப்பாடு அல்லது தனி ஒலியனியல் கட்டுப்பாடு இவற்றில் ஒன்றை உருப்படுத்தம் செய்யும் ஒருகுழும முற்றுநிலை மாற்றிகள் இடை மட்டம் மற்றும் புறமட்டம் இவற்றிற்கிடையில் பொருத்தமாற்றம் செய்ய இணையாகச் செயல்படுகின்றது. ஒவ்வொரு மட்டத்திற்கும் வரிவடிவ மற்றும் ஒலியனியல் உருப்படுத்தங்கள் இருக்கின்றன. உள்ளீடு சொல் வடிவில் இருக்கும் உரையிலிருந்து-பேச்சு பயன்பாடுகளுக்கு (எ.கா. ஒழுங்குமுறை சொல்லின்

சொல்சார் அடையாளம், அதன் சொல்வகைப்பாடு, அதன் திரிபு இவற்றை அறியும் உரை ஆக்கத்திற்கு) முற்றுநிலை மாற்றிகள் சொல் வடிவிலிருந்து புற உச்சரிப்புக்குப் பொருத்த மாற்றம் செய்ய இயலும். உள்ளீடு புற எழுத்துக்கூட்டலாக இருக்கும் உரையிலிருந்து-பேச்சுப் பயன்பாடுகளுக்கு (எ.கா. "உரையை உரக்கப் படிக்கும்" பயன்பாடுகளுக்கு) முற்றுநிலை மாற்றிகள் உள்ளுறையும் சொல் வடிவு வழி புற வரிவடிவத்திலிருந்து புற உச்சரிப்பை பொருத்தமாற்றம் செய்யவியலும்.

இறுதி சிக்கல் தலைப்பெழுத்தன்களைப் (acronyms) பற்றியதாகும். தலைப்பெழுத்தன்களைக் கால அளவுடன் அல்லது கால அளவில்லாமல் எழுத்துக்கூட்டலாம். கால அளவுடன் உள்ள தலைப்பெழுத்தன்களை எழுத்துக்கூட்டி உச்சரிக்கலாம் (எ.கா [ai rɛs]). கால அளவின்றி தோன்றும் தலைப்பெழுத்தன்களைச் (AIDS, ANSI, ASCAP) சொல்லாக எழுத்துக்கூட்டிவோ உச்சரிக்கவோ செய்யலாம்; எனவே AIDS எனதை aid என்ற வினையின் படர்க்கைஇட வடிவு போன்றே உச்சரிக்கலாம். லிபர்மான் மற்றும் சர்ச் (Liberman and Church, 1992) சொற்களாக உச்சரிக்கப்படும் தலைப்பெழுத்தன்களின் ஒரு சிறிய அகராதியை வைத்திருக்கவும் மீதமுள்ளவற்றை எழுத்துக்கூட்டவும் யோசனை கூறுகின்றார்.

2.2.5. உரையிலிருந்து-பேச்சில் மீக்கூறு

முன்னர் கூறிய வரிவடிவிலிருந்து ஒலிக்குச் செல்லும் மற்றிச் செயல்பாங்கு பேச்சை உருவாக்கும் உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் பகுதிக்கு உள்ளீட்டின் முக்கியக் கூறை உருவாக்கும. உள்ளீட்டின் மற்றொரு முக்கிய பாகம் மீக்கூறின் குறிப்பீடாகும். மீக்கூறு என்ற கலைச்சொல் பொதுவாகச் சொற்களஞ்சியத்திலிருந்து ஆக்கப்பட்ட ஒலிகளின் தொடர்ச்சியால் விளக்கப்படாத ஒரு வாக்கியத்தின் உச்சரிப்பின் நோக்குகளைக் குறிப்பிட பயன்படும். மீக்கூறு ஒலிகளைக் காட்டிலும்

நீண்ட மொழி அலகுகளின் மீது செயல்படும்; எனவே இது மீக்கூறு நடப்பின் ஆய்வு என்று சிலநேரங்களில் அழைக்கப்படும்.

மீக்கூறுக்கு மூன்று முக்கிய ஒலியனியல் நோக்குகள் இருக்கின்றன: மேன்மை (prominence), அமைப்பு (structure) மற்றும் சுரம் (tune). மேன்மை அசையழுத்தத்தையும் (stress) ஒலியழுத்தத்தையும் (accent) உள்ளடக்கும். மேன்மை என்பது அசைகளின் தன்மையாகும்; இது ஒரு அசை மற்றொரு அசையைக்காட்டிலும் மேன்மையுடையது என்று ஒத்தறி அடிப்படையில் விளக்கப்படும். உச்சரிப்புச் சொற்களஞ்சியங்கள் சொல்சார் அசையழுத்தத்தைக் குறியீடுசெய்யும்; எடுத்துக்காட்டாக table என்பது அசையழுத்தத்தை அதன் முதல் அசையிலும் machine என்பது அதன் அசையழுத்தத்தை இரண்டாவது அசையிலும் கொண்டிருக்கும். there, the அல்லது a போன்ற செயல்பாட்டுச் சொற்கள் பொதுவாக ஒலியசையைக் கொண்டிருக்காது. சொற்கள் ஒன்றாக இணைக்கப்படுகையில், அவற்றின் ஒலியசையழுத்த அமைப்பொழுங்கு இணைந்து முழு வாக்கியத்திற்கும் ஒரு பொரிய ஒலியசையழுத்த அமைப்பொழுங்கை உருவாக்கும். எவ்வாறு ஒலியசையழுத்தங்கள் இணைகின்றன என்பதில் முரண்பாடு இருக்கின்றது. ஒலியசைகயின் இடம் கருத்தாடல் காரணிகளால் மிக வலுவாகப் பாதிக்கப்படும்.

சில சொற்கள் இயல்பாகவே ஒன்றாகக் குழுமுவதாகத் தெரிகின்றது என்ற அளவில் வாக்கியங்களுக்கு மீக்கூறு அமைப்பு இருக்கின்றது; சில சொற்களுக்கு அவற்றிற்கிடையில் காணத்தக்க இடைவெளி அல்லது பிரிநிலை இருக்கின்றது. பெரும்பாலும் மீக்கூறு அமைப்பு மீக்கூறு தொடராக்கம் (prosodic phrasing) அடிப்படையில் விளக்கப்படும். இப்பெரிய மீக்கூறு அலகுகளுக்குப் பயன்படுத்தப்படும் கலைச்சொற்கள் இசையோட்டத் தொடர் (ntonational phrase or IP) இசையோட்ட அலகு (intonational unit) மற்றும் சுர அலகு (tone unit) என்பனவற்றை உள்ளடக்கும்.

2.2.5.1. மீக்கூறின் ஒலியல் அல்லது ஒலியியக்கவியல் நோக்குகள்

மூன்று ஒலியனியல் காரணிகள் ஊடாட்டம் செய்யும் மற்றும் அவை பல வேறுபட்ட ஒலியியல் மற்றும் ஒலியக்கவியல் நடப்புகளால் மெய்ப்படுத்தம் செய்யப்படும். மேன்மையான அசைகள் மேன்மையற்ற அசைகளைவிட பொதுவாக ஓசையுடையதாகவும் நீண்டதாகவும் இருக்கும். மீக்கூறுத் தொடர் எல்லைகள் பெரும்பாலும் விட்டிசையாலும் எல்லைக்கு முன்னர் அசையின் நீட்சியாலும் சில சந்தர்ப்பங்களில் எல்லையில் இசைமையின் குறைப்பாலும் தொடாப்படும். இசையோட்டச் சுரம் அடிப்படை நிகழ்வெண் (F0) வெளிவரம்பில் வெளிப்படுத்தப்படுகின்றது.

2.2.5.2. பேச்சு உருவாக்கத்தில் மீக்கூறு

உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் ஒரு முக்கியமான வேலை மீக்கூறின் பொருத்தமான மொழி உருப்படுத்தத்தை உருவாக்குவாதாகும்; அவற்றிலிருந்து வெளியீட்டு பேச்சுப் அலைவடிவில் வெளிப்படுத்தப்படும் பொருத்தமான ஒலியிக்க அமைப்பொழுங்கை உருவாக்குவதாகும். இத்தகைய மீக்கூறுப் பகுதியைக்கொண்ட ஒரு உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் வெளியீடு ஒரு கால அளவும் F0 (இசைமை) மதிப்பும் உள்ள ஒலிகளின் தொடச்சியாகும். ஒவ்வொரு ஒலியின் கால அளவு ஒலியியல் சூழலைப் பொறுத்து அமையும்.

2.2.6. தமிழில் உரையிலிருந்து பேச்சு

தமிழில் உரையிலிருந்து பேச்சு உருவாக்கத்திற்கான முன்னேற்பாடுகள் பல நடந்துள்ளன. குறிப்பாக சென்னையிலுள்ள IIT-யில் இது தொடர்பான ஆய்வுகள் மேற்கொள்ளப்பட்டுள்ளன. அந்நிறுவனத்தில் பணிபுரிந்த யக்ஞநாராயணா, இராசேந்திரன் போன்றோர் உரையிலிருந்து பேச்சுத்தொடர்பான கருவிகளைச் செய்வதற்கு முயற்சிகள் மேற்கொண்டு வெற்றிபெற்றுள்ளனர். அண்ணாப்பல்கலைக்கழகத்தில் நடைபெற்ற RCTDIL-T என்ற ஆய்வுத்திட்டத்திலும் இதற்கான முயற்சிகள் செய்யப்பட்டுள்ளன. திருவனந்தபுரத்திலுள்ள உலகத் திராவிட மொழிகளின் பள்ளி (ISDL) மலையாளத்திற்கு உருவாக்கிய மலையாள உரையிலிருந்து பேச்சு ஒழுங்குமுறை தமிழ் உரைக்கும் நல்ல பேச்சு வெளியீட்டைத் தருகின்றது.

2.3. எழுத்துத்திருத்தியும் இலக்கணப்பிழைத் திருத்தியும்

சொல்லாய்வின் வழி ஆவணங்கள் (documents) உருவாக்கப் படுகின்றன. ஆவணம் ஒரு கடிதமாகவோ, சிறு குறிப்பாகவோ, கட்டுரையாகவோ, நூலாகவோ இருக்கலாம். சொல்லாய்வியில் ஆவணத்தை உருவாக்க பொதுவாக பின்வரும் மூன்று முறைகள் பயன்படுத்தப்படுகின்றன:

1. கணிப்பொறியின் விசைப்பலகை (key board) வழி உள்ளீடு செய்தல்
2. பேச்சுணரி ஒழுங்குமுறையைப் (Voice Recognition System (VRS)) பயன்படுத்தி உரையை உள்ளீடு செய்தல்: இவ்வாறு செய்யும்பொழுது ஒலி வடிவில் பேசப்படும் செய்திகள் உரை வடிவிற்கு மாற்றப்படுகின்றன. இதற்குப் பேச்சிலிருந்து-உரை (Speech-to-Text (STT)) என்ற தொழில் நுட்பம் உதவுகிறது.
3. தட்டச்சு செய்தல் அல்லது அச்சடிக்கப்பட்ட உரையை ஒளி வழி எழுத்துணர்தல் (Optical Character Recongnition (OCR) என்ற தொழில்

நுட்பத்தின் மூலம் கணிப்பொறிக்கு உள்ளீடு செய்து சொல்லாய்விற்கு ஆவணமாக மாற்றுவது.

உரையை உள்ளீடு செய்ய மேற்கண்ட மூன்று முறைகளுள் ஏதேனும் ஒன்றைப் பின்பற்றலாம். பின்னர் சொல்லாய்வின் மூலம் அவைகளைக் கையாளலாம். உரைக்குள் எழுத்துப் பிழைகளும் இலக்கணப்பிழைகளும் இருப்பதற்கான வாய்ப்புகள் உள்ளன. ஆங்கில உரைகளில் உள்ள எழுத்துப்பிழைகளைக் கணிப்பொறி தானாகவே புரிந்துக்கொண்டு மாற்றுவதற்கு சொல்லாய்வில் வசதிகள் உள்ளன. இதற்கு எழுத்துப் பிழைகளைக் கண்டறிந்து திருத்துதல் என்று பெயர். இத்தகைய பிழைதிருத்தி தமிழ் சொல்லாய்வுகளில் உருவாக்கப்பட்டு வருகின்றன. எழுத்துப் பிழைகள் நிகழ்வதற்கான வாய்ப்புகள், எழுத்துப்பிழைகளின் வகைகள், பிழைகளைக் கண்டறிந்து திருத்துவதற்கான செயல்பாட்டு வழிமுறைகள் போன்றவற்றையெல்லாம் விரிவாக ஆராய்ந்து செயல்பட வேண்டி உள்ளது.

2.3.1. எழுத்துப் பிழை நிகழ்வதற்கான வாய்ப்புகள்

உரைச் செய்திகளைக் கீழ்கண்ட ஏதேனும் ஒருங்கு முறைகளில் உள்ளீடு செய்யப்படும்போது எழுத்துப்பிழைகள் நிகழ்வதற்கு வாய்ப்புகள் உள்ளன.

- விசைப்பலகை வழி உள்ளீடு செய்யும் பொழுது தட்டச்சுப் பிழைகள் (Typographical Errors) ஏற்படுகின்றன.
- பேச்சுணரி ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்ற பொழுது சொல்லின் உச்சரிப்பு மாற்றங்களினால் பிழைகள் ஏற்படுகின்றன. இத்தகைய பிழைகளுக்கு (Pronunciation Variation) என்று பெயர்.

- ஒளி எழுத்துணரி OCR வழி உள்ளீடு செய்கின்றபொழுது கருவிகளின் குறைபாடுகளினாலோ அல்லது எழுத்துக்களை சரியாக கருவிகள் படித்தறிதல் இல்லாததாலோ பிழைகள் தோன்றுகின்றன. உதாரணமாக ஆங்கிலத்தில் l என்ற எழுத்திற்கும் 1 என்ற எழுத்திற்கும் வேறுபாடு தெரியாததால் பிழை ஏற்படுகின்றன. பின்வரும் அட்டவணையில் சில சரியான எழுத்துக்களும் அவற்றிற்கிணையான தவறாகப் புரிந்துகொள்ளப்படும் எழுத்துக்களும் பட்டியலிடப்பட்டுள்ளன:

சரியன எழுத்து	தவறாகப் புரிந்துகொள்ளப்பட்ட எழுத்து
I	1
D	0
ri	n
0	∅

மேற்கண்டவாறு மூன்று வகையான உள்ளீட்டு முறைகளிலும் அதற்கான கருவிகளைப் பயன்படுத்தும் பொழுது பிழைகள் தோன்றுகின்றன. பேச்சறிதல் ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்ற பொழுது ஒரே வார்த்தையை அதன் இலக்கணக்கூறு மற்றும் தன்மைகளுக்கு ஏற்ப வெவ்வேறு விதமாக உச்சரிக்க வேண்டியுள்ளது. எடுத்துக்காட்டாக ஆங்கிலத்தில் object என்ற சொல் பெயர்ச்சொல்லாகவும் வினைச்சொல்லாகவும் பயன்படுகின்றது. பெயர்ச்சொல்லாகப் பயன்படும்போது ஆப்ஜிக்ட் [ˈɔbʤikt] என்றும் வினைச்சொல்லாகப் பயன்படும்போது அப்ஜெக்ட் [əbˈʤekt] என்றும் உச்சரிப்பு செய்யவேண்டியுள்ளது. அதைப்போல் the என்ற சொல்லைச் சூழலுக்கு ஏற்ப தி [ði:] என்றும் த [ðə] என்றும் மாறுப்பட்ட இருவகைகளில் உச்சரிக்க வேண்டியுள்ளது. இவ்வாறாக, ஒரே சொல்லிற்கான உச்சரிப்பு மாற்றங்களுக்கு திட்டவட்டமாக நன்கு வரையறுக்கப் பட்ட விதிமுறைகள் உள்ளன. அவ்விதிமுறைகளைக் கணிப்பொறிக்கு ஏற்ற வண்ணம் வடிவமைப்புகளாக வடிவமைத்து இத்தகைய பிழைகள் நீக்கப்படுகின்றன.

2.3.2. பிழை நீக்கும் வழிமுறைகள்

உரைச் செய்திகளில் உள்ள எழுத்துப் பிழைகளை கண்டறிந்து மாற்றுவதற்கு இருவழிமுறைகள் பயன்படுத்தப்படுகின்றன:

1) பெய்சியன் விதிகள் (Bayesian Rules),
2) ஓசை ஊடக மாதிரி (Noisy channel model).

பெய்சியன் விதிமுறைகளைப் பின்பற்றும்பொழுது எழுத்துப் பிழைகளைக் கண்டறிவதற்கான வாய்ப்புகளை ஆராய்வதற்கு கணிதமுறையில் அமைந்த நிகழ்தகவுக் கோட்பாடு (Probability Theory) பயன்படுத்தப்படுகிறது. பெரும்பாலும் இம்முறைகள் தட்டச்சு செய்த உரைகளுக்குப் பயன்படுத்தப்படுகின்றன. இம்முறையை ஒளிவழி எழுத்துணரி (OCR) முறையிலும் பேச்சுப் புரிந்துகொள்ளல் ஒழுங்குமுறைகளிலும் (Voice recoginition system) உள்ளீடு செய்யும் உரைகளுக்கும் பயன்படுத்தலாம்.

இரண்டாவது ஓசை ஊடக மாதிரி (Noisy Channel Model) முறையில் உள்ளீடு செய்யப்படுகின்ற உரைகள் எவ்வாறு மாற்றமடைகின்றன, மாற்றமடைவதற்கான காரணங்கள் யாவை போன்றவைகள் ஆராயப்பட்டு ஒரு மாதிரி வடிவம் (Model) தயாரிக்கப்படுகின்றன. பின்னர் இம்மாதிரி வடிவத்திற்கான எதிர்மறைச் செயல்பாடுகள் (Reverse process) ஒரு மாதிரிப் படிவமாக உருவாக்கப்பட்டு பிழைகள் நீக்கப்படுகின்றன. எடுத்துக்காட்டாக, தொலைபேசியில் பேசுகின்றபொழுதும் ஒலிவாங்கி வழி பேசுகின்றபொழுதும் ஒருவர் பேசும் குரல் அமைப்புகள் மாறுபடுகின்றன. இத்தகைய கருவிகள் மூலம் கிடைக்கப் பெறுகின்ற முடிவுகள் ஒலி அமைப்பைப் பொறுத்துள்ளது; உள்ளபடியாகவே அமையாது. இந்நிலையில்

அம்மாறுபாடுகள் நிகழ்வதற்கான காரணங்களை ஆராய்ந்து அதற்கு எதிர்மறையான திட்டங்களை வடிவமைக்கின்றபொழுது இவைகள் மூலம் நிகழும் பிழைகளை நீக்கலாம்.

2.3.3. எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் சிக்கல்களின் வகைகள்

எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் தொடர்பான சிக்கல்கள் பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

- மொழியில் இல்லாத சொற்களைக் கண்டறிதல் (Non-Word Error Detection)
- எழுத்துப்பிழை உள்ள செய்திகளை மட்டும் தனிமைப்படுத்தி ஆராய்தல்
- எழுத்துப்பிழை உள்ள சொல்லின் முன்பின் உள்ள அல்லது அச்சொல் இடம்பெற்ற தொடரின் பொருளை மையமாகக் கொண்டு பிழை நீக்கி சரிசெய்தல்

எழுத்துப்பிழைகளை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1. தட்டச்சுப்பிழைகள் (Typographical Errors)
2. சொல்லின் எழுத்து வடிவம் உரையாளருக்குத் தெரியாததால் ஏற்படும் பிழைகள்

தட்டச்சுப் பிழை பின்வரும் நான்கு வகைகளாகப் பிரிக்கப்படுகின்றன:

1. ஒரு எழுத்து விடுப்பட்டிருத்தல் (Delection Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் 'ட்' என்ற எழுத்தில்லாமல் 'சடம்' என இருப்பது விடுபடு பிழையாகும். இவ்வகையான பிழையை ஆங்கிலத்தில் 'Delection Error' என்பர்.

2. ஒரு எழுத்துக் கூடுதலாக இருத்தல் (Insertion Error): எடுத்துக்காட்டாக, *சட்டடம்* என்ற பிழையானச் சொல்லில் 'ட' என்ற எழுத்துக் கூடுதலாக இருப்பது செருகல் பிழையாகும். இவ்வகையான பிழைகளை ஆங்கிலத்தில் 'Insertion Error' எனபர்.
3. ஒரு எழுத்திற்குப் பதிலாக வேறொரு எழுத்தை மாற்றி அமைத்து விடுதல் (Substitution Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் முதல் எழுத்து 'ச' தவறாக அடிக்கப்பட்டும்போது அத்தவறான எழுத்து 'ச' ஆகவோ 'க' ஆகவோ 'ப' ஆகவோ இருக்கலாம். இம்மாதிரியான பதிலீட்டுப் பிழைகளை ஆங்கிலத்தில் (Substitution Error) என்பர்.
4. அடுத்தடுத்த இரண்டு எழுத்துக்கள் இடம் மாற்றி இருப்பது (Transposition Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் பிழையாக 'ட' என்ற எழுத்தும் 'ட்' என்ற எழுத்தும் இடம் மாறி *சடட்ம்* என்று மாற்றி அடிக்கப்படுவது இடமாற்றத் தவறாகும். இவ்வகையானப் பிழைகளை ஆங்கிலத்தில் 'Transposition Error' என்பர்.

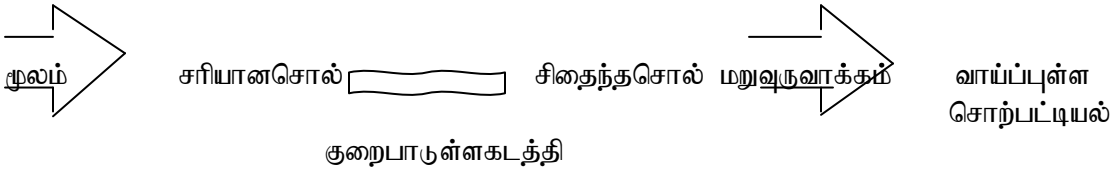
2.3.4. மொழியில் இல்லாதச் சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல்திட்டம்

ஒரு மொழியில் உள்ள சொல்லினைக் கண்டறிந்து சரிசெய்வதற்கு அகராதி பயன்படுத்தப்படுகிறது. கொடுக்கப்பட்ட சொல் அகராதியில் இல்லை எனில் அச்சொல்லை மொழியில் இல்லாத சொல் என கருதலாம். இதற்குத் தகுந்த அறிவுரையைக் கணிப்பொறி பயன்பாட்டாளருக்கு வழங்கும். இதற்காகப் பயன்படுத்தப்படும் கணிப்பொறிக்கான அகராதி சிறிதாக இருத்தல் நலம். மிகப்பெரிய அகராதியாக இருந்தால் அதிகமாகப் புழக்கத்தில் உள்ள

எழுத்துப்பிழைச் சொற்கள் கூட அரிதாகப் பயன்படுத்தும் சரியான சொல்லாக இருக்கும்.

2.3.4.1. எழுத்துப்பிழைத் திருத்தி (Bayesian Medhod)

ஒரு உரையிலுள்ள எழுத்துப்பிழையைக் கண்டறிந்து திருத்துவதற்குப் பல செயல்திட்ட முறைகளை வடிவமைக்கப்படுகின்றன. அவற்றில் பெய்சியன் முறை நிகழ்தகவு அடிப்படையில் (Probability Thories) அமைந்ததாகும். இம்முறையின் அடிப்படைத் தத்துவம் பின்வரும் படத்தில் விளக்கப்பட்டுள்ளது:



உரைக்கான சொற்கள் தகுந்த உள்ளீட்டுக் கருவிகள் மூலம் உட்படுத்தப்படுகின்றன. உள்ளீடுக் கருவி கணிப்பொறியின் விசைப்பலகையாகவோ அல்லது ஒலியுணரி ஒழுங்குமுறையின் ஒலிவாங்கியாகவோ இருக்கலாம். தட்டச்சுச் செய்யும்பொழுது எழுத்துக்களின் சிதைவோ அல்லது ஒலியுணரி ஒழுங்குமுறையில் ஒலிச் செய்தியின் சிதைவோ ஏற்பட வாய்ப்புகள் உள்ளன. மூலக்கருவிகள் மூலம் உருவாக்கப்படும் சொல் உரிய கடத்திகள் (Channels) மூலம் கடத்தப்படுகிறது. கடத்திகளில் உள்ள குறைபாடுகள் காரணமாகவும் கடத்தும் செய்திகளில் சிதைவு ஏற்படலாம். இவ்வாறாகச் சிதைந்த சொல் மறு உருவாக்கம் (Decode) செய்யப்படுகிறது. இச்சொல் மூலக்கருவி மூலம் உருவாக்கப்பட்ட சொல்லுக்கு வாய்ப்புள்ள சொல்லாகவே அமையும். இவ்வாறு சிதைந்த சொல்லுக்கு வாய்ப்புள்ள சொற்களின் பட்டியல்

உருவாக்கப்படும். இப்பட்டியலிலிருந்து சிதைந்த சொல்கூறு திருத்தம் செய்யப்பெற்று பிழை நீக்கிய சரியானச் சொல் அடையாளம் காணப்படுகின்றது.

பேசியன் முறையில் எழுத்துத் திருத்திக்கான வழிமுறைகள் இருபெரும் கட்டங்களாக செய்யப்படுகின்றன:

1. பிழை நீங்கிச் சரியான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டடியல் தயாரித்தல்
2. இச்சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லுக்கும் மதிப்பெண் அளித்து அதிக மதிப்பெண் பெறும் சொல்லைப் பிழைத் திருத்தம் செய்யப்பட்ட சொல்லாகக் கருத வழிவகை செய்தல்

இம்முறையில் பிழை திருத்தம் என்பது சொல்லில் ஏதேனும் ஒரு எழுத்து மட்டும் பிழையாக இருக்கலாம் என்ற வாதத்தின் அடிப்படையில் செயல்முறைகள் மேற்கொள்ளப்படுகின்றன. எனவே பின்வரும் வகையிலான எழுத்து பிழைகள் மட்டும் இம்முறையில் எடுத்துக் கொள்ளப்படுகின்றன:

- சொல்லின் ஒரு எழுத்து விடுபட்டு இருத்தல்
- சொல்லில் ஒரு எழுத்து அதிகமாக இருத்தல்
- சொல்லில் ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து தவறாக இருத்தல்
- சொல்லில் அடுத்தடுத்த இரு எழுத்துக்கள் இடம் மாறி இருத்தல்

பேசியன் முறையில் எழுத்துத் திருத்தி உருவாக்குவதற்கு முதலில் பெருந்தரவு (Corpus) ஒன்றினைக் கணிப்பொறியில் உருவாக்கிக் கொள்ள வேண்டும். இந்த பெருந்தரவில் மொத்தம் ‘N’ சொற்கள் இருப்பதாகக் கெர்ளவோம். பிழையானச் சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் ஒன்றை உருவாக்க வேண்டும். இப்பட்டியலை

உருவாக்குகின்றபொழுது ஒரு சொல்லில் ஒரு எழுத்தினால் மட்டும் பிழை நிகழ்வதாக எடுத்துக் கொள்ளப்படுகிறது. எனவே சொல்லில் (அ) ஒரு எழுத்து விடுபட்டு இருத்தல், (ஆ) ஒரு எழுத்து கூடுதலாக இருத்தல், (இ) ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருத்தல், (ஈ) அடுத்தடுத்த இரு எழுத்துக்கள் தவறாக இடம் மாறியிருத்தல் ஆகிய நான்கு மட்டும் கருத்தில் கொள்ளப்படுகின்றன. மேற்கண்ட முறைகளின்படி பிழையான சொல்லிற்கு வாய்ப்புள்ள சொற்பட்டியலில் 'V' சொற்கள் இருப்பதாகக் கொள்வோம். இப்பட்டியலில் 'c' என்பதை ஏதாவது ஒரு சொல்லாகக் கொள்வோம். 'c(c)' என்பது 'c' என்ற சொல் பெருந்தரவில் எத்தனை முறை இடம்பெற்றிருக்கிறது என்பதைக் குறிப்பதாகக் கொள்வோம். இப்போது பிழை நீக்கிய சொல் 'c' என்ற சொல்லாக இருப்பதற்கான முன்நிகழ்வு எண்ணிக்கை பின்வரும் சூத்திரத்தின் மூலம் கணக்கிடப்படுகிறது:

$$P(c) = \frac{c(c)+0.5}{N+0.5V} - 1$$

$$P(\text{அடம்}) = \frac{c(\text{அடம்})+0.5}{N+0.5V}$$

(v = வாய்புள்ள சொல், C = விரிதரவு (corpus)

மேற்கண்ட சூத்திரம் ஒரு சொல் மட்டும் பெருந்தரவில் இடம்பெற்றுள்ள முன்நிகழ்வு (Prior) மதிப்பைத் தருகிறது. இரண்டாவது கட்டநிலையில் தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்களுக்கு மதிப்புகள் கண்டறியப்படுகின்றன. இதற்கு மாதிரி உரைகள் ஒரு தட்டச்சர் அல்லது பல தட்டச்சரைக் கொண்டு உள்ளீடு செய்யப்படுகின்றன. இவ்வுரையை அடிப்படையாகக் கொண்டு நான்கு விதமாக அணிகள் (Matrices) உருவாக்கப்படுகின்றன. தவறாகத் தட்டச்சு செய்த சொல் ஆங்கிலத்தில் 'typographical error' எனப்படுகிறது.

- del (x,y) என்பது xy என்ற இரு எழுத்துகளுக்குப் பதிலாக x மட்டும் எத்தனை முறை தவறாக அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
- ins (x,y) என்பது x என்ற எழுத்துக்களுக்குப் பதிலாக xy என்று தட்டச்சு செய்துள்ள எண்ணிக்கையைக் குறிக்கிறது.
- sub (x,y) என்பது x என்ற எழுத்திற்குப் பதிலாக எத்தனை முறை y என்ற எழுத்து தட்டச்சு செய்யப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
- trans (x,y) என்ற எழுத்துக்களுக்குப் பதிலாக yx என இடமாறி அச்சடிக்கப் பட்டுள்ளது என்பதைக் குறிக்கிறது.

தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் இருந்து ஏதாவது ஒரு சொல் 'c' என்பதை எடுத்துக்கொள்வோம். 'c_p' என்பது 'c' என்ற சொல்லில் p-இடத்தில் உள்ள எழுத்தைக் குறிப்பதாகக் கொள்வோம். இந்நிலையில் 'c' என்ற சொல் வாய்ப்பான சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

1. ஒரு எழுத்து விடப்பட்டிருக்கும் நிலையில்

$$P(t/c) = \frac{del[c_{p-1}, c_p]}{count\ [c_p - 1c_p]}$$

2. ஒரு எழுத்து கூடுதலாக இருக்கும் நிலையில்

$$P(t/c) = \frac{ins\ [c_p - 1, t_p]}{count\ [c_{p-1}]}$$

3. ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருக்கும் நிலையில்

$$P(t/c) = \frac{trans\ [t_P, c_p]}{count\ [c_p]}$$

4. இரு எழுத்துக்கள் இடம் மாறி இருக்கும் நிலையில்

$$P(t/c) = \frac{trans[c_p,c_p+1]}{count[c_p,c_p+1]}$$

தவறான சொல் வாய்ப்புள்ள சொற்பட்டியலில் உள்ள 'c' என்ற சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

$$W= P(t/c)*P(c)-3$$

இந்த சூத்தீரத்தில் P(c) என்பது c என்ற சொல் பெருந்தரவில் உள்ள முன் நிகழ்வு மதிப்பைத் தருகிறது. P(t/c) என்ற தவறாக தட்டச்சு செய்யப்பட்டு 'typo' என்ற சொல் 'c' என்ற சொல்லாக இருப்பதற்கு எவ்வளவு வாய்ப்புகள் உள்ளன என்ற மதிப்பைத் தருகின்றது. இவ்விரு எண்களின் பெருக்கல் தொகை, அதாவது P(t/c)□□□*P(c) என்பது இச்சொல்லிற்கான மதிப்பெண்ணைத் தருகிறது.

மேற்கண்டவாறு தவறான சொலலுக்கு வாய்ப்புள்ள சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லிற்கும் மேற்கண்ட மதிப்பினைக் கண்டறிய வேண்டும். எந்த செயால் அதிக மதிப்பெண் பெற்றிருக்கிறதோ அந்த சொல் அதிக சரியான சொல் என கருதப்படுகிறது. அதிக மதிப்பெண் பெற்றிருக்கும் சொல் பின்வரும் சூத்திரத்தின் மூலம் கூறப்பட்டுள்ளது.

$$W=\underset{c\epsilon V}{argmax}\ \frac{P(t/c)P(c)}{P(t)}$$

மேற்கண்டவாறு அதிக மதிப்பெண் பெற்ற சொல் இடம் பெற்ற வாக்கியத்தின் பொருளுடன் ஒத்துப் பார்க்கப்படும். அது வார்த்தையின் பொருளுக்கேற்ற சொல் இல்லையெனில் அடுத்த மதிப்பெண் பெற்ற சொல் பொருத்திப்பார்க்கப்படும். இவ்வாறாகத் தவறாகத் தட்டச்சு செய்யப்பட்டுள்ள typo சரியான சொற்களாக மாற்றச் செயல் திட்டங்கள் உள்ளன.

இவ்வாறாக உரைகள் எழுதுதம்போது பிழைகளைக் கண்டறிந்து திருத்துவதற்குக் கணிப்பொறியில் பல செயல்திட்டங்கள் உருவாக்கப்படுகின்றன. அதில் மிகவும் பிரபலமானது மேற்கண்ட பேசியன் முறையாகும். இதுபோல் குறைந்த அளவு வித்தியாசம் உள்ள (minimum distance) என்ற மற்றொரு முறையும் பயன்படுகிறது.

2.4. உருபனியல் ஆய்வு

தமிழ்மொழி ஒரு திராவிட மொழி. இம்மொழி வினைச்சொல்லை இறுதியாகக் கொண்ட வாக்கியங்களை உடையதாகும். இம்மொழி இலக்கணச் செயற்பாங்குகளை (அதாவது இலக்கணப் பொருண்மைகளை) உருபனியலில் நிலையில் வெளப்படுத்தும் ஒட்டுக்கள் நிறைந்த மொழியாக விளங்குகிறது. இம்மொழியில் உள்ள ஒவ்வொரு வேர் உருபனும் ஆயிரக்கணக்கான சொற்களின் உருவாக்கத்திற்கு (அதாவது திரிபு மற்றும் ஆக்க வடிவுகளின் உருவாக்கத்திற்கு) அடிப்படையாக அமைகிறது. தமிழ் வாக்கியங்கள் எழுவாய்தொடரையும் வினையையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் செயப்படுபொருள் தொடரையும் வினையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் பெயர்த்தொடரையும் (அதாவது பயனிலைத் தொடரையும்) கொண்ட அமைப்பாகவோ வரும். இதன் காரணமாகத் தமிழ் மொழி ஒரு SOV மொழி என்பர். தமிழ் மொழி ஒரு முன் மாதிரி SOV மொழிக்குள்ள எல்லாப் பண்புக்கூறுகளையும் பெற்றுள்ளது. பெயர்கள் வேற்றுமை உருபுகளுக்காகத்

திரிபுறுகின்றன. பெயர்த்தொடர்களுக்கும் வினைகளுக்கும் உள்ள உறவுகள் வேற்றுமை உருபுகளாலோ பின்னுருபுகளாலோ வெளிப்படுத்தப்படுகின்றன. வினைகள் கால இடைநிலைகளை ஏற்பதாகவும் எழுவாத்தொடருக்கு உடன்பாடாக வினைமுற்று விகுதிகளை ஏற்பதாகவும் வருகின்றன. இதன் காரணமாக எந்தக் கணிப்பொறி தொடர்பான மொழி ஆய்விற்கும் உருபனியல் பகுப்பாய்வு இன்றியமையாததாகவும் அடிப்படையாகவும் அமைகின்றது. மேலும் எழுத்துப்பிழை திருத்தி உருவாக்கத்திற்கும் இலக்கணப்பிழை திருத்தி உருவாக்கத்திற்கும் உருபனியல் பகுப்பாய்வு அடிப்படையாகும்.

உருபனியல் ஆய்வியன் உருவாக்கம் எந்த இயற்கை மொழி ஆய்வின் செயல்பாட்டிற்கும் மிக முக்கியமான முதற்படியாகும். வேறுபட்ட மொழி வகைகளுக்கு வேறுபட்ட ஆய்வு நெறிமுறைகள் இருக்கின்றன. எளிய இணைப்பு உருபனியல் உள்ள (simple concatinative morphology) மொழிகள் சொல்லும் செயற்பாங்கும் (item and process) மற்றும் சொல்லும் வரிசைமுறையும் (item and arrangement) என்ற மொழி மாதிரிகளால் உருவாக்கப்பட்ட ஆய்விகளால் தமது தேவையைப் பூர்த்தி செய்துகொண்டுள்ளன. பண்புக்கூறுகள் அடிப்படையிலான ஒருமைப்படுத்தும் கோட்பாடுகள் (unifivation theories) ஸ்பானிஷ் போன்ற மொழிகளுக்குப் பயன்படுத்தப்பட்டது. இக்காலத்தைய மிகப் புகழ்வாய்ந்த கணினிசார் இலக்கணமான தலைமையால் இயக்கப்படும் தொடரமைப்பு இலக்கணம் (Head-Driven Phrase Structure Grammar (HDPSG) பண்புக்கூறு அடிப்படையிலான உருபனியலைப் (Feature Basecd Morphology) பயன்படுத்துகின்றது. இங்கு தமிழ்மொழியின் உருபனியல் ஆய்வுக்குப் பயன்படுத்தப்பட்ட மாதிரிகளும் வேறு முன்னணி மாதிரிகளும் விளக்கப்படும்.

கணிப்பொறி வழி ஆய்வைப் பொறுத்தவரையில் இருமட்ட மாதிரி (two level model) உருபனியலின் முக்கிய கோட்பாடுகளில் ஒன்றாகும். இது பரந்த பரப்பெல்லையில்படும் மொழிகளுக்குச் செயல்படுத்தப்பட்டிருக்கின்றது. முற்றுநிலை தானியங்கி (Finite State Automata (FSA)) மாதிரி சென்னையிலுள்ள AUKBC ஆய்வு மையத்தால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வில் (API Morphological Analyzer) பயன்படுத்தப்பட்டுள்ளது. தலைமைத் தொடரில் அமைப்பு இலக்கணம் (Head Phrase Structure Grammar (HPSC)) தமிழுக்கு உமாராணியால் (Umarani, 2001) பயன்படுத்தப்பட்டுள்ளது. ஆம்பிள் என்ற மற்றொரு மாதிரி சொல்லும் வரிசையும் என்ற மாதிரியுடன் தொடர்புடையது.

பகுத்துலும் உருவாக்கலும் (Parsing and Generation) இயற்கை மொழிகளின் உருபனியல் ஆய்வின் இரு முக்கியமான நேர்வுகள் ஆகும். பகுப்பாய்வு என்பது பகுப்பாய்வி தரப்பட்ட உள்ளீட்டை புரிந்து கொள்ளவும் ஆயவும் செய்து பெரும்பாலும் அவற்றைப் பொருண்மை மற்றும் இலக்கணப் பண்புக் கூறுகளுடன் வெளியீடாகத் திருப்பித்தரும் செயல்பாடாகும். பல உரை மீட்பு ஒழுங்குமுறைகளுக்கும் (Text Retrieval Systems (TRS)) இயந்திர மொழிபெயர்ப்புச் (Machine Translation (MT)) செயல்பாடுகளுக்கும் பகுப்பாய்வு முக்கியமானது ஆகும். அதாவது இந்த இரண்டு செயல்பாடுகளுக்கும் தொடக்க இடம் பகுப்பாய்வாகும். ஒரு உணரி (recognizer) உள்ளீட்டுச் சொல்லை வேறுபட்ட உருபன்களாகப் பகுப்பாய்வு செய்கின்றது. ஆங்கிலச் சொல்லான reads என்பது பின்வருமாறு புரிந்து கொள்ளப்படும்: read என்ற வினையில் படர்க்கை இட ஒருமை எண் குறிப்பிடப்பட்ட நிகழ்கால வடிவம். பகுப்பாய்வி உணரியிலிருந்து அது புரிந்து கொள்ளும் சொற்களுக்குப் பொருத்தமான அமைப்பைத் தருகிறது என்ற அளவில் வேறுபடுகின்றது. இவ்வாறு ஒரு உருபனியல் பகுப்பாய்வி உள்ளீடு செய்யப்பட்ட சொல்லை அதன் உறுப்பு உருபன்களாகப் பிரித்து அதனைப் படிநிலை அமைப்பாகப் பகுப்பாய்வு செய்கிறது. இது சொல்லாக்கத்தில்

உருபன்களை இணைப்பதற்கு மொழியில் குறிப்பிட்ட இயக்க நெறியின் ஒரு நோக்கை வெளிப்படுத்துகிறது. இவ்விக்கநெறி ஒரு வேருடன் கட்டுண்ட உருபுகளை முன்னொட்டாகவோ உள்ளொட்டாகவோ பின்னொட்டாகவோ வேருடன் சேர்க்கும். தமிழ்மொழியில் இது கூடுதலாகப் பின்னொட்டாக்கம் ஆகும். முன்னொட்டுக்கள் சமஸ்கிருதத்தில் இருந்து வந்த சில கடன் சொற்களில் இருக்கின்றன. மாறாக, உருபனியல் உருவாக்கி (Morphological Generator) உருபன்களை இணைத்துச் சொற்களை உருவாக்கும் நோக்கைப் பயன்படுத்துகிறது. இவ்வாறு உருபனியல் உருவாக்கி பகுப்பாய்வதன் மறுதலையாகும். இதன் உள்ளீடு ஒரு குழும உருபன்களோ அதன் இலக்கண அமைப்புகளோ ஆகும். இது அவற்றை ஒன்றிணைத்து ஒரு மொழியில் நல்வடிவாக்கமுள்ள சொற்களை உருவாக்குகிறது. உருபனியல் பகுப்பாய்வி இச்செயல்பாடுகளில் ஒன்றையோ அதற்கு மேலானவற்றையோ செய்கின்றது.

வேறுபட்ட உருபனியல் ஆய்விகள் அவை விளக்கும் மொழிகளின் அடிப்படையில் வேறுபட்ட கோட்பாடு அணுகுமுறைகளைப் பயன்படுத்துகிறது. அதாவது வேறுபட்ட மொழிகளின் இயல்பும் கலவைத்தன்மையும் அவற்றை ஆயப்பொருத்தமான கோட்பாடு அணுகுமுறைகளைத் தீர்மானிக்கிறது. வேறுபட்ட சட்டங்களின் அடிப்படையில் வேறுபட்ட உருபனியல் ஆய்விகள் உருவாக்கப்படுகின்றன. வேறுபட்ட சட்டகங்களின் தேர்வும் அவற்றின் வேறுபட்ட நடைமுறைபடுத்தல்களும் ஒப்பீடு செய்யும் ஆய்வுக்கு அவசியமாகிறது.

2.4.1. உருபனியல் ஆய்விகளும் கோட்பாட்டு அணுகுமுறையும்

இப்பகுதியில் இருமட்ட உருபனியலைப் பயன்படுத்தும் பிசிகிம்மோ (PCKimmo), ஆம்பிள் (Ample) என்பனவும் சொல்லும் வரிசைமுறையும் என்ற மாதிரியின் அடிப்படையில் அமைந்த ஜிஎஸ்மார்ஃபும் (GSMorph) சொல்லும் செயற்பாங்கும்

மாதிரியின் அடிப்படையில் அமைந்த இராமசாமியின் (Ramaswamy, 2000) தமிழ் உருவாக்கியும் (Generator for Tamil) சொல் ஒலியனியல் (Lexical Phonology) அணுகுமுறை அடிப்படையில் அமைந்த இரங்கநாதனின் (Renganathan,1997) டேக் தமிழும் (Tag Tamil) சொல்லும் வரிசைமுறையும் மற்றம் சொல்லும் செயற்பாங்கும் என்ற இரு மாதிரிகளையும் பயன்படுத்தும் வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும் இராசேந்திரனின் உருபனியல் ஆய்வியும் விளக்கப்படுகின்றன. இவை தவிர தமிழில் முயற்சிக்கப்பட்டுள்ள பிற உருபனியல் ஆய்விகளும் பட்டியலிடப்பட்டுள்ளன.

2.4.2. தமிழில் உருபனியல் ஆய்விகள்

தமிழில் கணிப்பொறி தொடர்பான செயல்பாடுகளுக்கு அடிப்படையாக அமைவது உருபனியல் பகுப்பாய்வும் உருவாக்கமும் ஆகும். பலர் இதில் பங்களிப்பு செய்துள்ளனர். பல உருபனியல் ஆய்விகள் தமிழில் உருவாக்கப்பட்டுள்ளன. இவற்றில் முக்கியமானவைகளைப்பற்றிய சிறு குறிப்பு கீழே தரப்பட்டுள்ளன.

2.4.2.1. இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி

முதல் முதலில் இந்திய மொழிகளுக்கிடையில் இயந்திர மொழி பெயர்ப்பை மேற்கொள்ள வேண்டி அனுசாராகா (anusaraka 'language acessor') என்ற மொழிபெயர்புக் கருவி உருவாக்கத்திற்காகப் பாணினி வடிவவாதம் அடிப்படையில் ஆய்வு மேற்கொண்ட குழுவைச் சார்ந்த ஆய்வாளர்களால் DOE நிதி நல்கையில் தொடங்கிய ஆய்வுதான் இந்தியாவில் இயற்கை மொழி ஆய்விற்கு வித்திட்டது எனலாம். இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்பு ஆய்வுக்காக உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்விகள் அடிப்படையில் தமிழ்-இந்தி மொழி

பெயர்ப்புக்காகத் தமிழ் உருபனியல் பகுப்பாய்வியும் இராசேந்திரன் உதவியுன் உருவாக்கப்பட்டது. AUKBC ஆய்வு நிறுவனக் குழுவினரால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வியும் இராசேந்திரனின் மேற்பார்வையில் உருவாக்கப்பட்டதுதான். தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் பல்கலைக்கழக மானியக் குழுவின் நிதிநல்கையில் கீழ் மேற்கொள்ளப்பட்ட “Spell and Grammar Checher for Tamil” என்ற உயராய்வுக்காக PROLOG வழியமைப்பு மொழியைப் பயன்படுத்தி ஒரு உருபனியல் பகுப்பாய்வி இராசேந்திரனால் உருவாக்கப்பட்டது. இப்பகுப்பாய்வி இந்திய மொழிகளின் நடுவண் நிறுவன விரிதரவில் பயன்படுத்தப்பட்டு மதிப்பீடு செய்தபோது 95% வெற்றிகரமாக பகுத்துக்குறித்தல் வெளியீடு கிடைத்தது.

2.4.2.2. கணேசனின் உருபனியல் பகுப்பாய்வி

கணேசன் (Ganesan, 1994, 2003) அவர்கள் தமிழுக்காக இந்திய மொழிகளின் மையநிறுவனம் உருவாக்கிய மூன்று மில்லியன் சொற்கள் அடங்கிய விரிதரவை (CIIL Corpus) ஆய வேண்டி ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இவர் ஒலியனியல் மற்றும் உருபனியல் விதிகளைக் கையாண்டு இப்பகுப்பாய்வியை உருவாக்கினார். சமீபத்தில் சிறந்த தமிழ் உருபனியல் பகுப்பாய்வியை உருவாக்கியுள்ளார்.

2.4.2.3. கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி

கபிலன் (Kabilan, 1994) அவர்கள் தமிழ் வினைகளுக்கான ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இது பேராசிரியர் தெய்வசுந்தரத்தின் (மொழியியல்

பகுதி, தமிழ்மொழித் துறை, சென்னைப் பல்கலைக்கழகம்) மேற்பார்வையின் கீழ் நடந்த தமிழ்சார் இயற்கை மொழி ஆய்வின் விளைவாகும்.

2.4.2.4. தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்

தெய்வசுந்தரம் தமிழுக்கென ஒரு சொல்லாய்வியைத் (Word Processor) தயாரித்துள்ளார். இது மிக நன்றாகச் செயல்பாடு செய்கின்றது. இது இருநிலை மாதிரியை (two-level-model) அடிப்படையாகக் கொண்டது. புணர்ச்சி விதிகள் பகுப்பாய்வுக்குப் பயன்படுத்தப்படுகின்றன. இவர் ஆய்வி (Deiva sundaram and Gopal 2003) உருபொலியனியல் மற்றும் உருபனியல் என்ற இருநிலைகளில் செயல்படுகின்றது. இது ஒலியன் சேர்ப்பையும் (phonemic addition), ஒலியன் மாற்றத்தையும் (phonemic change) கணக்கிலெடுக்கின்றது. இவ்வாய்வி உருபனியல் அமைப்பொழுங்கு (mophotactics) அடிப்படையில் ஒட்டுகள் பகுதியுடன் இணைக்கப்படும் நிரலைக் கணக்கிலெடுத்துக் கொண்டு பகுப்பாய்வு செய்யும்.

2.4.2.5. AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி

AUKRC ஆய்வு நிறுவனம் (தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையுடன் இணைந்து) உருவாக்கிய API உருபனியல் சொல்லாய்வி தரவுக்கோப்பு, நிலை மாற்றமைவு அட்டவணை (State table.dat) என்பனவற்றில் தந்துள்ள தானியங்கி அட்டவணையைச் சார்ந்து இருக்கிறது. API உருபனியல் சொல்லாய்வி சொற்களை அதன் திரிபுகளிலிருந்து பிரித்தெடுக்கும்படி வடிவமைக்கப்பட்டுள்ளது.

2.4.2.6. வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்

வைஷ்ணவி (Vaishnavi Ramaswamy, 2003) முனைவர் பட்ட ஆய்வுக்காகத் தமிழ் உருபனியல் பகுப்பாய்வை எடுத்துக்கொண்டு ஆய்வேட்டினைச் செய்துள்ளார். மேலும் வைஷ்ணவி (Vaishnavi Ramaswamy, 2002) தமது ஆய்வியல் நிறைஞர் ஆய்வேட்டிற்காக தமிழுக்கு ஒரு உருபனியல் உருவாக்கியை உருவாக்கி விளக்கமாகத் தந்துள்ளார். அவருடைய உருபனியல் உருவாக்கி சொல்லும் வரிசைமுறையும் மாதிரியையும் சொல்லும் செயற்பாங்கும் மாதிரியையும் பயன்படுத்தி செய்யப்பட்டுள்ளது. இந்த உருவாக்கி பிசிகிமோவின் (Pckimmo) கூட்டிணைப்பு நெறிமுறையால் செயல்படுகின்றது. வைஷ்ணவி இதைக் கலப்பு மாதிரி (hybrid model) என்று அழைக்கின்றார்.

PERL–இல் உருவாக்கப்பட்ட மற்றும் செயல்படுத்தப்பட்ட கலப்பு மாதிரி (hybrid model) உள்ளுருவாக்கப்பட்ட சொற்களஞ்சியத்தின் சட்டகத்தையும் மொழியின் வேறுபட்ட ஒவ்வொரு சொல் வகுப்பிற்கும் சொல்லும் வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் என்பனவற்றின் ஒரு குழும விதிகளையும் கொண்டிருக்கிறது. தரப்பட்ட உருபொலியனியல் விதிகளின் அடிப்படையில் செயல்படும் இணையான துணை நடைமுறைகளை (corresponding subroutines) வேண்டுவதன் மூலம் இதன் ஆய்வு நிருபிக்கப்படுகிறது. இது ஒவ்வொரு வகைப்பாட்டின் சொல் வடிவுகளைக் கையாளும்படி வடிவமைக்கப் பட்டுள்ளது. இது முதன்மை அமைப்பின் கோப்புகளால் கட்டளையிட்ட படி செயல்படுகிறது. வழியமைப்பு மொழியில் மொழித் தகவலின் மரபுரிமை குறியாக்கம் இருக்கிறது. திறந்த இறுதியுள்ள உருபொலியனியல் மற்றும் மாற்றுருபு விதிகளின் (open ended morphophonimic and allomorphic rules) வழி மொழித் தகவலின் பகுப்பாய்வால் இது நடைப்பெறுகிறது (Vaishnavi Ramaswamy, 2003:101-110).

இம்மாதிரி உருபனியல் சொல்லும்வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் மாதிரிகளின் கலவையில் அதன் கோட்பாட்டு அடிப்படையை அமைத்துக் கொள்கிறது. செயற்பாட்டைச் சொற்பகுதியில் (word stem) உருபனியல் இயல்பைப் பொறுத்து ஒரு மாற்றத்தை (transformation) விவரணை செய்வதாக விளக்க இயலும். திரிபு நீக்கத்தின் (inflectional stripping) ஒவ்வவொரு நிலையிலும் குறிப்பிடப்பட்டுள்ள கட்டுப்பாடுகளின் முழுத்தொடர்சிகளின் வழி கடந்த பின்னர் ஒரு சொல் பகுப்பாய்வு செய்யப்படுகிறது.

ஒரு பகுப்பாய்வுத் தொகுதி (Analysis Module) வலமிருந்து இடமாகக் கோர்வை அமைவதை வருடிச் (scanning) செய்கிறது. ஒரு நேரத்தில் ஒரு பின்னொட்டை வருடி அறிந்து பிரித்தெடுக்கிறது. மற்றும் ஒவ்வொரு நேர்விலும் எடுத்துக்காட்டப்பட்ட ஒலியனியல் மற்றும் உருபொலியனியல் விதிகளின் உதவியுடன் மீதிச்சொல்லை மீட்டுருவாக்கம் செய்கிறது. கோர்வை தீர்வது வரை இது நடைபெறும். வெளியீடு கண்டுபிடிக்கப்பட்ட உருபன்களின் வடிவில் இருக்கும். பின்னர் அவை உறுதி செய்யப்படும்.

2.4.2.7. வின்ஸ்டன் குருசின் ஜி.எஸ். மார்ஃப் பகுப்பாய்வியும் உருவாக்கியும்

ஜி.எஸ். மார்ஃப் (GS Morph) என்று சுருக்கமாகக் குறிப்பிடபப்படும் எழுத்து அடிப்படையிலான அமைப்பு மொழியியலார் உருபனியல் பகுப்பாய்வி (Grapheme Oriented Structuralist Morphological Processor (GS Morph)) (Sengupta 1999) ஆம்பிள் போன்று சொல்லும் வரிசை முறையும் மாதிரியின் அடிப்படையில் அமைந்தது. இது ஒவ்வொரு உருபனின் விரிவான புற மெய்ப்படுத்தத்தை வேண்டும் அல்லது ஏற்படுத்தும். ஆம்பிள் ஒட்டுக்களை அவை பின்னொட்டுக்களா உள்ளொட்டுக்களா என்பதன் அடிப்படையில் வேறுப்பட்ட கோப்புகளில் சேகரிக்க இயலும். ஆனால்

ஜிஎஸ் மார்.்பின் பகுப்பாய்வு ஒரு வேர்ச் சொற்களஞ்சியத்தையும் (Root Lexicon) பின்னொட்டுக்களின் சொற்களஞ்சியத்தையும் (Suffix Lexicion) உட்படுத்தும்.

ஜிஎஸ் மார்.்ப் சொல்லடைவைப் பொருத்துவதால் செயல்படுகிறது. அதன் உள்ளீடு எழுத்து உருப்படுத்தத்தின் வடிவில் இருக்கிறது. அது வலது எல்லையிலிருந்து இடது எல்லைக்குப் பகுப்பாய்வைத் தொடங்குகிறது. முதலாவது அது தரப்பட்ட உள்ளீட்டைப் பூஜிய பின்னொட்டால் தொடரப்படும் வேராகக் கருதுகிறது. இதன்படி இது பின்னொட்டுச் சொற்களஞ்சியத்தில் பின்னொட்டுக்காகப் பரிசோதிக்கும். பின்னர் அது வேர்க்கோப்பில் பூஜிய ஒட்டிற்கு நேராக இருக்கும் சொல்லடைவுகளைப் (concordance) பார்க்கும். அது உள்ளீட்டைப் பொருத்துவதில் தோல்வியுற்றால் அது சொல் வடிவத்தைப் பற்றிய அதன் ஆரம்ப அனுமானத்தை மீளப்பெறும். அது உள்ளீட்டின் எழுத்துக்களை வெட்டி எடுக்கத் தொடங்கும். ஒவ்வொரு தடவையும் ஒரு எழுத்தை வெட்டி எடுக்கும்போது அது எல்லா வெட்டி எடுக்கப்பட்ட எழுத்துக்களும் சேர்ந்த ஒன்றைப் பின்னொட்டுச் சொற்களஞ்சித்தில் பொருத்தத்திற்காகப் பார்க்கும். வெட்டி எடுக்கும் செயற்பாங்கு முழுகோர்வையும் தீருவது வரை தொடர்ந்து நடைபெறும். பின்னொட்டுச் சொற்களஞ்சியத்தில் தேடல் வெற்றி பெற்றால் உள்ளீட்டின் மீதிப்பகுதி வேராக முன்மொழியப்பட்டு வேர்க்கோப்பில் பார்க்கப்படும். முன்மொழியப்பட்ட வேர் சொற்களஞ்சியத்தில் காணப்பட்டால் பின்னொட்டு மற்றும் வேர்க் கோப்புகளில் இருந்து அவற்றின் தொடர்புடைய பொருள் தரப்படும். இது வேர் மற்றும் பின்னொட்டுக் கோப்புகளைச் சொல்லடைவு செய்வதன் மூலமும் செய்யப்படும். சொல்லடைவு செய்யப்படும் சமயத்தில் வேர் மற்றும் பின்னொட்டுச் சொற்களஞ்சியங்களுக்குச் சிறு அட்டவணைகள் (hash tables) உருவாக்கப்படும். வழிமுறை வரைவுகள் (algorithms) தொடர்புடைய சிறு அட்டவணைகளில் தொடர்புடைய சொல்லடைவுகளைப் பொருத்திப் பார்க்க வேண்டி பரிசோதிக்கும். முழு உள்ளீடும் தீர்க்கப்படுவது வரை இச்செயற்பாங்கு நிற்பதில்லை என்பதை முக்கியமாகக் கவனிக்க வேண்டும். இப்படி அது ஒரு பொருத்தத்தைக் கண்டுபிடித்து வெளியீட்டைத் தந்தாலும், சாத்தியம்

என்றால் பகுப்பாய்வி கூடுதலான பகுப்புகளைக் கண்டு பிடிக்க இயலும் என்பதாகும். இது வேர் மற்றும் மாற்றுருபுகளுக்கு ஒரே ஒலியனியல் வடிவு இருந்தால் பகுப்பாய்வு இரண்டிற்கும் எல்லாச் சாத்தியமான பகுப்பாய்வையும் தரும். தற்போதைய நடைமுறைப்படுத்தலில் விளைவு (effective) வேறுபாட்டைக் காலப் பின்னொட்டில் பார்க்கப்படும் நிலையில் தமிழில் வினைகளுக்கு ஒரு தனி வேர் உள்ளீட்டைத் தருவது, ஜி.எஸ். மார்.்பில் வேர் மற்றும் பின்னொட்டுக் கோப்புகள் அமைக்கப்படும் முறையின் காரணமாக இது சாத்தியமாகிறது. இது பகுப்பாய்வின் போது ஒரே நடப்பை மிக எளிதாகப் பொருண்மை மயக்கம் தீர்க்க உதவுகிறது.

ஜி.எஸ் மார்.்ப் பகுப்பாய்வி சொல்லின் வரிவடிவ உருபடுத்தத்தை உள்ளீடாக ஏற்கின்றது. சொற்கள் இடது எல்லையிலிருந்து பகுப்பாயப்படுகின்றன. முன்னர் விளக்கப் பட்டதுபோல் வேர் மற்றும் பின்னொட்டு சொற்களஞ்சியங்களின் பயன்பாட்டால் பகுப்பாய்வு வெற்றிகரமாக நடைபெறும்.

ஜி.எஸ். மார்.்பின் உருவாக்கி உட்கூறு பகுப்பாய்வுக்கான அதே சொற்களஞ்சியங்களைப் பயன்படுத்துகின்றது. உள்ளீடு தரப்படும்போது உருவாக்கி அதை வேர்ச் சொற்களஞ்சியத்தில் பரிசோதித்து அதற்குத் தரப்பட்ட சொல்லடைவை எடுக்கும். அது பின்னர் அதே சொல்லடைவைக் கொண்ட கோர்வைகளைப் பின்னொட்டு அகராதியில் தேடும். ஏதாவது கோர்வை அதே சொல்லடைவுடன் இருந்தால் வேர்மாற்றுரு அப்பின்னொட்டுக் கோர்வையுடன் சேரும். இது பின்னொட்டில் உள்ள உருபன்களின் அர்த்தங்களுடன் திருப்பி அனுப்பப்படும். பகுப்பாய்வுக்குப் பயன்படுத்தப்படும் அதே கோப்புகள் பயன்படுத்தப்படுவதால் உருவாக்குதலும் துல்லியமாக இருக்கின்றது. (Winston Cruz, 2002:127).

வின்ஸ்டன் குருஸ் (Winston Cruz, 2000) தமிழ் வினைச்சொற்களைப் பகுப்பாய்வு செய்வதற்கு ஜி.எஸ். மார்.ப் முறையைக் கையாண்டார். ஜி.எஸ் மார்.ப் என்பது உருபன் அமைப்பொழுங்கு குறியீட்டால் ஆனது. வழிமுறை வரைவு (Algorithm) இரண்டு கோப்புகளை பொருத்தமானதா இல்லையா என்று மட்டுமே ஆய்வு செய்கிறது. சொல்லாய்வி இரண்டு வடிவங்களை மட்டுமே பயன்படுத்தி இரண்டு கோப்புகள் படைக்கப்பட்டுள்ளது.

2.4.2.8. துரைபாண்டியின் உருபனியல் பகுப்பாய்வு

தற்கால தமிழில் வினைச்சொல் வடிவத்தினை உருவாக்கும் பகுப்பாய்வு இயந்திரத்தை முழுமையாக உருவாக்கியுள்ளார்.

2.4.2.9. RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி

RCILTS-T (Resource Centre for Indian Language Technological Solutions-Tamil) என்ற அமைப்பின் தமிழ் உருபனியல் பகுப்பாய்வி அட்சரம் (atcharam) என அழைக்கப்படுகின்து. அட்சரம் வேர்ச் சொல்லிலிருந்து ஆக்கச் சொல்லைத் தனியாகப் பகுப்பாய்வு செய்து உருபனுடன் தொடர்புபடுத்துகிறது. இதனைப் பயன்படுத்தி 15 வகைகளை அடிப்படையாகக் கொண்டு இருபதாயிரம் வேர்ச்சொற்கள் தமிழ் அகராதியில் உருவாக்கப்பட்டுள்ளன. இது இரண்டு தொகுதிகளை (modules) உடையது. அதாவது பெயர் மற்றும் வினைச் சொல்லை அடிப்படையாகக் கொண்டது. இது ஊக விதிகளுடன் (heuristic rules) பொருண்மை மயக்கம் உடையதாகும். அதற்கு வினை மற்றும் பெயர் திரிபுகளைக் கையாண்டுள்ளனர்.

2.4.2.10. RCILTS-T –இன் உருபனியல் உருவாக்கி

இந்நிறுவனம் தமிழ்மொழியலும் உருபனியல் உருவாக்கியைத் தயார் செய்துள்ளது. அதற்கு அட்சயம் (Atchayam) எனப் பெயரிடப்பட்டுள்ளது. அட்சயம் என்பது பெயர் மற்றும் வினைச்சொற்களைக் கொண்டு இரண்டு முக்கிய உள்ளீடுகளைப் படைத்துள்ளது. இதில் பெயர்ச்சொல் பிரிவில் பின்னொட்டு இடம்பெறும். எடுத்துக்காட்டாக, பன்மை ஒட்டுக்களையும் திரிபு வடிவங்களையும் வேற்றுமை உருபுகளையும் பின்னுருபுகளையும் உருவாக்கியுள்ளனர். இதே போல் வினைப்பிரிவில் காலங்களையும் இடம்-எண்-பால் உருவாக்கிகளையும் பெயரச்சம் (relative), வினையெச்சம், பின்னொட்டு மற்றும் துணை வினைச்சொற்களையும் கொண்டுள்ளது. இது சந்தி விதிகள் மற்றும் 125 உருபனியல் விதிகளையும் பயன்படுத்தி உள்ளது. இது பெயரடை மற்றும் வினையடையையும் கையாண்டுள்ளது. இதில் சொல் மற்றும் வாக்கிய உருவாக்கியைக் கொண்டு ஆய்வு செய்யப்பட்டுள்ளது.

2.4.2.11. பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்

சொல் வடிவில் பொருண்மை மயக்கம் இருக்கும். இதற்கு மாறாக ஒப்புருச் சொல்லின் (homonymy) வகைப்பாடுகளில் எதிர் பாராத குறுக்கீடுகள் உருவாவதால் உருபனியல் பிறழ்ச்சி, ஒட்டுக்களின் பல்வகைச் செயல்பாடுகள் அல்லது நிச்சயமற்ற பின்னொட்டுச் சொற்கட்டுகளையும் (word bounding) உருவாக்குகிறது. இவ்விதம் வாக்கிய அமைப்புச் சூழல்களை இயல்பாக தீர்மானிக்கின்ற பகுப்பாய்வு ஆகும். இதனையே பொருண்மை மயக்கம் அல்லாத உருபனியல் என்கிறோம். இரண்டு அடிப்படை அணுகுமுறைகள் சமவாய்ப்பு (probability) உடையது. விதியை அடிப்படையாகக் கொண்ட அணுகுமுறையில் சில பொருண்மை மயக்கம் தீர்வு காணப்பட முடியாமல் உள்ளது. ஆனால் மிகச் சில தவறுகளைக் கொண்டுள்ளது.

பொதுவாகப் புள்ளியியல் அடையாளப்படுத்திகள் முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளைத் தருகின்றன. ஆனால் அவற்றறுள் நிறைய தவறுகள் காணப்படுகின்றன. சமவாய்ப்புள்ள முறையில் 1980-இருந்து உருபனியலில் பொருண்மை மயக்கமற்ற பகுப்பாய்வே அதிகம் இடம் பிடித்துள்ளது. நிலையான புள்ளியியல் முறைகளின் பயன்பாடு முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளாக உள்ளது. 'Influence of Morphology in Word Sense Disambiguation for Tamil' (தமிழ் உருபனியலில் பொருண்மை மயக்கம் அல்லாத சொல் செல்வாக்கு உடையது). பரிசோதனைகளை மேற்கொள்ளும்பொழுது மேற்பார்வை மற்றும் பகுதி மேற்பார்வை அணுகுமுறையை மேற்கொண்டு உருபனியல் திரிபுகளின் பாதிப்புமுறைகளை பற்றியும் திடமான தீர்வான உருபனியலில் பொருண்மை மயக்கம்மற்ற பகுப்பாய்வை தமிழ்மொழியிலும் பிற திரிபு மொழிகளிலும் பயன்படுத்தலாம் எனத் தீர்வு கண்டுள்ளனர்.

2.5. கணினி தொடரியல்

கணினி தொடரியல் ஆய்வு வாக்கியங்களைத் தொடர்களாகவும் சொற்களாகவும் பகுத்து அவற்றின் தொடரியல் மற்றும் சொல்வகைப்பாட்டுப் பண்புக்கூறுகளை அடையாளப்படுத்தித் தரும் செயல்பாடாகும். வாக்கியங்கள் தரப்படுகையில் அவற்றை தொடரியல் கூறுகளாகப் பகுப்பது மற்றும் தொடரியல் கூறுகள் தரப்படும் போது வாக்கியங்களை உருவாக்குவது என்ற இரு செயல்பாடுகள் இதில் அடங்கும்.

2.5.1. கணிப்பெறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்

ஒரு கூற்றின் பொருத்தமான பொருளைத் தருமாறு சொற்களுக்கு இடையே உள்ள உறவுகளை பிரித்தெடுக்கும் நுட்பத்தைத் தருவது இந்த இலக்கண

வடிவமைப்பின் முக்கியமான செயல்பாடாகும். வாக்கியங்களை பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் எனப்படும். முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது. அவ்வாக்கியங்களுக்கு அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல். தொடக்ககால கட்டத்தில் சூழல்வரையறை இல்லாத இலக்கண வடிவவாதம் பயன்படுத்தப்பட்டது. தற்போது கட்டுப்பாடு அடிப்படையிலான இலக்கண முன்மாதிரிகள் பயன்படுத்தப்படுகின்றன. அவை கீழே பட்டியலிடப்பட்டுள்ளன:

- செயல்பாட்டு ஒருங்கிணைக்கப்பட்ட இலக்கணம் (Functional Unification Grammar)
- தலை இயக்கத் தொடரமைப்பு இலக்கணம் (Head-driven phrase Structure Grmmar)
- செயல்பாட்டு இலக்கணம் (Fuctional Grammar)
- சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar)
- வகைப்பாட்டு ஒன்றிணைக்கப்பட்ட இலக்கணம் (Categorical Unification Grammar)
- கிளை சேர்க்கப்படும் இலக்கணம் (Tree adjoining Grammar)

2.5.2. ஆழமில்லாப் பகுப்பாய்வு

நாம் ஆழமில்லாப் பகுப்பாய்வு என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளியீட்டைக் காட்டிலும் நிறைவுடைய ஆய்வைக்

குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகின்றோம். ஒரு ஆழமில்லாப் பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அக அமைப்பையும் செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். ஆழமில்லாப் பகுப்பாய்வு மூன்று முன்மாதிரித் தொகுப்புகளை உள்ளடக்கியது:

1. சொல்வகை அடையாளப்படுத்தல் (parts-of-speech tagging): ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை) உறுதி செய்யும்
2. தொடர்பகுதியாகப் பிரித்தல் (chunking): இதில் சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்பும் தரப்படுகையில் எந்தச் சொற்கள் தொடர் பகுதிகளாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பித்தொடர் போன்றவைகளாக) குழுமப்பட வேண்டும் என்று இச்செயல்பாட்டால் தீர்மானிக்கப்படுகின்றது.
3. உறவைக் கண்டுபிடித்தல் (relation finding): ஒரு வாக்கியத்தில் உள்ள தொடர்பகுதிகள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனியலை, இடம் போன்றவை) தீர்மானிக்கப்படவேண்டும்.

2.5.3. தமிழில் சொல்வகை அடையாளப்படுத்திகள்

தமிழில் சொல் வகை அடையாளப்படுத்திகள் உருவாக்கப்பட்டுள்ளன. பின்வருவன அவற்றில் முக்கியமானவைகளாகும்:

1. வாசு அரங்கநாதனின் டேக் தமிழ்
2. கணேசனின் சொல்வகை அடையாளப்படுத்தி
3. RCILTS-தமிழின் கதம்பம்

2.5.4. தமிழில் தொடரியல் பகுத்துக் குறிப்பான்கள்

தமிழில் தொடரியல் பகுத்துக்குறிப்பான்கள் உருவாக்கும் முயற்சிகள் பல மேற்கொள்ளப்பட்டன. அவற்றில் சில இங்கு பட்டியலிடப்பட்டுள்ளன:

2.5.4.1. பாஸ்கரனின் தொடரியல் பகுத்துக்குறிப்பான்

முற்றுநிலைத் தானியங்கி மொழியமைப்பு மட்டத்தில் பகுத்துக்குறித்தலுக்கு முக்கியமான தொழில் நுட்பங்களில் ஒன்றாகும். பாஸ்கரன் (Baskaran, 1984) தமிழ் வாக்கியங்களில் பகுத்துக்குறித்தலுக்கு வேண்டி முற்றுநிலை தானியங்கையைப் பயன்படுத்தினார். இது தமிழ்ப்பல்கலைகழகத்தில் உருவாக்கப்பட்ட ருசியன்-தமிழ் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு வேண்டி இவரால் உருவாக்கபட்டது.

2.5.4.2. குமார சண்முகத்தின் தமிழ்த்தொடரியல் பகுத்துக்குறிப்பான்

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக 'Parse Representaton of Tamil Syntax' என்ற தலைப்பில் ஒரு ஆய்வேட்டைச் சமர்பித்துள்ளார். இவ்வாய்வேட்டில் இவர் தமிழ்த் தொடரியல் பகுப்பாய்வு பற்றி வரிவான விளக்கம் தந்துள்ளார். சார்பு இலக்கண வடிவவாதம் அடிப்படையில் இவர் தமிழுக்கு ஒரு தொடரியல் பகுத்துக்குறின்பானையும் வடிவமைத்துள்ளார்.

2.6. தமிழில் மொழிநடை ஆய்வி

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர், சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி மூலம் அறிய இயலும். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்பு பண்புக்கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த உரைகளை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்துக் கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனம் கண்டுக் கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய கணினி வழி தமிழ்மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியியல் மொழியியல் (Statistical linguistics), அளவீட்டு மொழியியல் (Quantitative linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணைப்புரிகின்றன. மொழி நடையை ஆய உதவும் மென்பொருள்கள் பல கண்டுப்பிடிக்கப்பட்டு பயன்பாட்டில் உள்ளன.

2.7. கணினிப் பொருண்மையியல்

பெருண்மையில் என்பது மொழிக் கூற்றுகளின் பொருண்மையை ஆய்வதாகும். நாம் இங்கு பொருண்மையைக் கண்டுபிடிக்க முறையான உருப்படுத்தங்களையும் கூற்றுக்களிலிருந்து இப்பொருண்மை உருப்படுத்தங்களுக்கு வேண்டி வழிமுறை வரைவுகளிலும் கவனக்குவிப்பு செய்யவேண்டும். நாம் எவ்வாறு கூற்றுக்களின் பொருண்மை தொடர்களின், சொற்களின் மற்றும் உருபன்களின் பொருண்மைகளுடன் தொடர்புடையவை என்று அறிந்திருக்கவேண்டும். பேசுபவர்கள் கேட்பவர்கள் மற்றும்

கூற்றுக்கள் காணப்படும் சூழல்கள் இவற்றுடன் தொடர்புடைய சிக்கல்கள் பயன்வழியியலில் தீர்மானிக்கப்படும்.

2.7.1. பொருண்மையை உருப்படுத்தம் செய்தல்

மொழிக் கூற்றுக்களின் பொருண்மை முறையான அமைப்புகளில் வெளிப்படுத்தப்படும். இதை மொழிகளின் பொருண்மை உருப்படுத்தம் என்பர். இப்பொருண்மை உருப்படுத்தங்கள் ஒலியனியல், உருபனியல் மற்றும் தொடரியல் உருப்படுத்தங்களைப் போன்று பங்களிப்புச் செய்கின்றது.

2.7.2. பொருண்மையியல் ஆய்வு

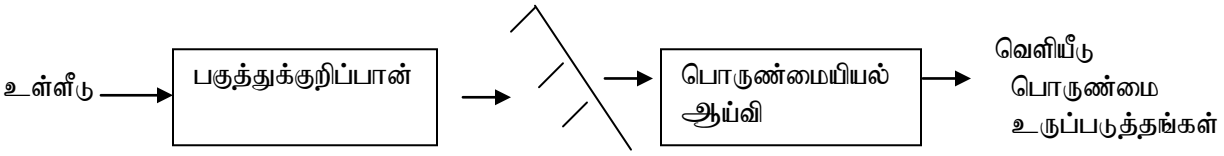
இங்கு பொருண்மை ஆய்வின் சிக்கல்களுக்கு கணினி அணுகுமுறை கருத்தில் கொள்ளப்படும். முதல் அணுகுமுறை தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வாகும் (Syntax-Driven Semantic Analysis). இதற்குக் குறைந்த அளவான பயன்பாடுதான் இருக்கின்றது. இது சொற்களஞ்சியத்திலிருந்தும் இலக்கணத்திலிருந்தும் நிலையான அறிவின் அடிப்படையில் உள்ளீடுகளுக்குப் பொருண்மை உருப்படுத்தங்களைத் தரும்.

2.7.3. தொடரியலால் இயக்கப்படும் பொருண்மையியல்

இது இணைப்பாக்கத்தின் கொள்கை அடிப்படையில் (principle of compostitionality) அமைந்தது. இவ்வணுகுமுறையில் உள்ளுறையும் முக்கிய கருத்து ஒரு வாக்கியத்தின் பொருண்மை அதன் பகுதிகளின் பொருண்மைகளால் இணைக்கப்பட்டது என்பதாகும். வாக்கியத்தின் பொருண்மை முழுவதும் அதில்

அடங்கும் சொற்களின் அடிப்படையில் அமையாமல் வாக்கியத்தில் வரும் சொற்களின் நிரல், குழுமுதல் மற்றும் சொற்களுக்கிடையேயுள்ள தொடர்புகள் இவற்றின் அடிப்படையில் அமையும். அதாவது வாக்கியத்தின் பொருண்மை அதன் தொடரியல் அமைப்பின் அடிப்படையில் அமையும் என்பதாகும். எனவே தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வில் பொருண்மை உருப்படுத்தங்களின் இணைப்பு ஒருவகை இலக்கணத்தால் தரப்படும் தொடரியல் கூறுகளாலும் உறவுகளாலும் வழிக்காட்டப்படும்.

உள்ளீடு செய்யப்படும் வாக்கியத்தின் தொடரியல் ஆய்வு பொருண்மையியல் ஆய்வின் உள்ளீடாக அமையும். கீழேத் தரப்பட்டுள்ளப்படம் இக்கருதலில் இருந்து நேரடியாகப் பெறப்படும் பைப்லைன் அடிப்படையிலான அணுகுமுறையை (pipeline-oriented approach) எடுத்துக்காட்டும்:



இவ்வாய்வு பொருண்மை ஆய்வின் உள்ளீடாகத் தரப்பட்டு பொருண்மை உருப்படுத்தத்தை உருவாக்கும்.

2.7.4. சூழல் கட்டுப்பாடில்லாத விதிகளில் பொருண்மையின் விரிவாக்கங்கள்

இக்கட்டுப்பாடில்லாத இலக்கண விதிகள் பொருண்மையியல் இணைப்புகளால் விரிவுப்படுத்தப்படும். இவ்விணைப்புகளை ஒரு கட்டுமானத்தின் உறுப்புகளின் பொருண்மையிலிருந்து அதன் பொருண்மை உருப்படுத்தம் கணிக்கப்படுகிறது

என்பதைக் குறிப்பிடும் கற்றல் விதிகளாக கருத இயலும். விரிதாக்கப்பட்ட விதிகளுக்குப் பின்வரும் அமைப்பு இருக்கும்.

2.7.5. முந்தையப்பகுத்துக்குறிப்பானின் பொருண்மையியல் ஆய்வை ஒருங்கிணைத்தல்

முந்தைய பத்திகளில் நாம் பொருண்மையியல் ஆய்விகளின் எளிய பைப் லைன் கட்டமைப்பைப் பார்த்தோம். இதில் முழு தொடரியல் பகுத்துக்குறித்தலின் கிளைகள் பொருண்மையியல் ஆய்வுக்கு செலுத்தப்படுகிறது. இதன் நோக்கம் என்னவென்றால் இணைப்பாக்க அணுகுமுறை அது தொடர்புப்படுத்தும் தொடரியல் பகுத்துக்குறித்தலை வேண்டுகின்றது என்ற கருத்துச்சாயலிலிருந்து வந்ததாகும். இருப்பினும் பொருண்மையியல் ஆய்வை தொடரியல் ஆய்வுக்கு இணையாகச் செய்ய இயலும். முந்தையப் பகுத்துக்குறித்தலுடன் பொருண்மை ஆய்வின் இணைப்பு நேரிடையானதாகும். இணைப்பாக்கக் கொள்கை மரபுத் தொடர்களின் (idioms) பயன்பாட்டால் சிக்கலை நேரிடும். மரபுத்தொடர்கள் பிரித்து பொருண்மையை ஆய இயலாத மொழி அலகுகளாகும். இது இணைப்புக் கொள்கைக்கு நேரிடையான விளைவைத்தரும்.

2.7.6. சொற்பொருண்மையியல்

சொற்களஞ்சியம் (அகராதி) சொற்களின் பொருண்மை அமைப்பு மற்றும் அவைகளுக்கிடையே உள்ள உறவுகள் எவ்வாறு அமைந்துள்ளன என்பதை விளக்கும். இந்த ஒழுங்குமுறையான பொருண்மையியலில் தொடர்புடைய அமைப்பைப்பற்றி படிப்பது சொற்பொருண்மையியல் ஆகும். சொற்பொருண்மையியல் சொற்களுக்கிடையேயுள்ள உறவுகளை தொடர் உறவுகளாகவும் அடுக்குறவுகளாகவும் ஆயும். சொற்களுக்கிடையேயுள்ள அடுக்குறவுகள் ஒருபொருள்பன்மொழி, உள்ளடங்கு

மொழி, உள்ளடக்கு மொழி, எதிர்மொழி போன்றவையாகும். மின்சொற்களஞ்சியங்களும் சொல்வலையும் சொற்களுக்கிடையேயுள்ள சொல்லுறவுகள் மற்றும் பொருண்மையுறவுகள் அடிப்படையில் அமைந்தவை.

2.8. தமிழில் சொற்களஞ்சியம் தயாரித்தல்

சொற்களஞ்சியம் (Thesaurus) தயாரிப்பதற்கு கணிப்பொறியைப் பயன்படுத்தலாம். ஒத்தல் (similarity), உள்ளடங்கல் (inclusion), மேலுறல் (overlapping), வேறுபடல் (disjunction) என்ற தர்க்கப்பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை அவற்றின் படிநிலை அமைப்பும் பொருண்மையும் வெளிப்படும் விதத்தில் பொருண்மைக் களங்களாகப் பகுத்துத் தருவது சொற்களஞ்சியம் ஆகும். கணிப்பொறி உதவியுடன் இச்சொற்களஞ்சியத்தை குறைந்த உழைப்பில் குறுகிய கால கட்டத்தில் தயாரித்து விடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்ததில் மின் சொற்களஞ்சியத்தை உருவாக்கலாம். இத்தகைய மின்சொற்களஞ்சியம் மொழிப்பெயர்த்தல், மொழிக் கற்பித்தல் போன்ற செயல்பாடுகளுக்கு பெரிதும் பயன்படும். தமிழுக்கான மின்சொற்களஞ்சியம் தமிழ்ப்பல்கலைக்கழக மொழியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து தயாரித்துள்ளன.

தரவு மையத்தில் சொற்கள் பொருண்மை அடிப்படையில் பகுப்பட்டு பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை வரிசைப்படுத்தித் தரவேண்டும். பின்வரும் எடுத்துக்காட்டு இதை விளக்கும்.

எ.கா.

1. பருமைப்பெயர்கள்

1.1 அ.ஃறிணைப்பெயர்கள்

1.1.1.உயிரில்லாதவை

1.1.1.1 இயற்கைப்பொருட்கள்

1.1.1.1.1விண்வெளிப்பொருட்கள்

1.1.1.1.1.1 வானம், ஆகாயம், வான், ஆகாசம், விண், விசும்பு, அந்தரிட்சம், அந்தரம்

1.1.1.1.1.2 கீழ்வானம், மேல்வானம்

இது தரவு மையமாக அமையும் இதிலிருந்து வழியமைப்புகள் வழிச் சொற்களஞ்சியத் தகவல்களைப் பெறலாம்.

2.9. தமிழில் சொல்வலை தயாரித்தல்

இது ஒரு இணையத் தளதரவு மையமாகும் (On line data base). இது சொற்களஞ்சியம் மற்றும் அகராதி என்ற இரண்டின் பண்புகளையும் கொண்டு மிகுந்த திறனுடையதாய் அமைந்துள்ளது. சொற்கள் தமக்கிடையே உள்ள பொருண்மை உறவுகளால் வலைபோல் பின்னப்பட்டிருப்பதை வெளிப்படுத்துவதால் இது அமையும். சொல் வலையின் மிகப் பரந்த நோக்கம் என்னவென்றால் சொல் தகவல்களை சொல் வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குப் படுத்துவதாகும். இதன் அடிப்படையில் சொல் வலை அகராதியிலிருந்து வேறுப்பட்டு சொற்களஞ்சியத்தையும் உள்ளடக்கும். ரொஜஸட் தெசாரஸ் போன் சொற்களஞ்சியங்களும், தமிழுக்கு இராஜேந்திரனால் (2001) உருவாக்கப்பட்ட தமிழ்ச்சொற்களஞ்சியமும் சொல் வலைப் பின்னுவதற்கு முன்னோடியாய் அமையும். இவையெல்லாம் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் அவற்றின் தேவைக்கு அதிகமான/மிகையான பதிவுகள் காணப்பெறும். இருசொற்கள், ஒருபொருள்

பன்மொழிகள் என்றால் கூட அகரவரிக்காக இரண்டும் இருமுறைப் பதிவுப்பெறும். தலைப்பு ஒழுங்கு முறையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் இரண்டு முறை நோக்கீடு செய்து தான் தேவையான தகவல்களை பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எப்பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்து கொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவல்களைப் பெற இயலுகின்றது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். ஆனால் கணிப்பொறியின் தேடலை விரைவாக செய்ய இயலும். சொல் வலை மின்சொற்களஞ்சியத்தை விட ஒருபடி மேலானது. தமிழ்ப்பல்கலைக் கழகமும், சென்னையிலுள்ள AUKBC Research Centrer-உம் இணைந்து தமிழ்சொல் வலையைத் தமிழ் இணையப் பல்கலைக்கழகத்திற்காக உருவாக்கியுள்ளது. இது முதல் கட்டநிலைதான். பொருளுதவி இல்லாத காரணத்தால் அடுத்தக் கட்டத்திற்குச் செல்ல இயலவில்லை.

2.10. தமிழில் சொற்களை அகரவரிசைப் படுத்துதல்

தமிழ்மொழியின் அகரவரிசைத் தொடர்பான செய்திகளையும் வழியமைப்பு ஆணைகளையும் (Programer Instructions) கணிப்பொறிக்குத் தந்தால் கணிப்பொறி மிகக் குறைந்த நேரத்தில் வேண்டிய சொற்களை அகரவரிசைப் படுத்தித் தந்துவிடும். ஆங்கிலத்திற்கு இது மிகச் சாதாரண ஒன்று. தமிழைப் பொருத்தவரையில் பலவிதமான எழுத்துருக்களும் (fonts) அதற்குப் பின்னணியாய் வேறுபட்ட குறியன்களும் (code) குறியனெண்களும் (code numbers) பயன்படுத்தப்படுவதால் அகரவரிசைப் படுத்தல் எளிய செயல்களாக அமைய வில்லை. இருப்பினும் வேறுப்பட்ட எழுத்துருக்களை வழியமைப்புகள் மூலம் இணைத்து அகரவரிசைப்படுத்தும் மென்பொருள்கள் கண்டுப்பிடிக்கப் பட்டு பயன்பாட்டில் உள்ளன.

2.11. தமிழில் சொல்லடைவு தயாரித்தல்

சொல்லடைவுத் தயாரித்தலையும் கணிப்பொறி மூலம் விரைவில் செய்துவிட இயலும். சொற்களையும் அவற்றோடு தொடர்பான செய்திகளையும் தந்தால் குறைந்த காலநேரத்தில் கணிப்பொறியின் உதவியால் சொல்லடைவைத் தயாரித்து விடலாம். எடுத்துக்காட்டாக, சங்க இலக்கியங்களுக்கு சொல்லடைவுத் தயாரிப்பதை எடுத்துக் கொள்வோம். பின்வரும் அடிகள் சங்க இலக்கியமான புறநானூற்றிலிருந்து எடுக்கப்பட்டது.

1. இவனியாரென்குவை யாயி னிவனே (புறம் 83.1)
2. தூங்குகையா னோங்கு நடைய (புறம் 22.1)

இவ்வடிகள் சொற்களாகப் பின்வருமாறுப் பிரித்துத் தரப்படவேண்டும். வழியமைப்புகள் மூலம் இச்சொற்களை அகரவரிசைப் படுத்தி அவை ஒவ்வொன்றும் வரும் இடங்களையும் குறிப்பிட்டுத் தர இயலும்.

ஆயின்	(புறம் 83.1)
இவன்	(புறம் 83.1)
என்குவை	(புறம் 83.1)
ஏ	(புறம் 83.1)
ஓங்கு	(புறம் 22.1)
கையால்	(புறம் 22.1)
தூங்கு	(புறம் 22.1)
நடைய	(புறம் 22.1)

ஒரு சொல்/சொல்லன் சங்க இலக்கியத்தில் எத்தனை முறை கையாளப்பெறுகின்றது. அல்லது குறிப்பிட்ட ஒரு நூலில் எத்தனை முறை கையாளப் பெறுகின்றது என்பன போன்ற தகவல்களை வழியமைப்புகள் மூலம் பெறலாம். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும், கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்தி சங்க இலக்கியங்களுக்கான ஒரு சொல்லடைவுத் தயாரித்துள்ளன.

2.12. தமிழில் சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவுத் தயாரித்தல்

சொல்லாட்சி அகராதி (concordance) ஒரு பனுவலில் இடம்பெறும் சொற்களை அகரவரிசைப் படுத்தித் தருவதோடு அச்சொற்கள் நூலில் இடம்பெறும் இடங்கள் எத்தனைமுறை இடம்பெறுகின்றன என்ற கணக்கீடு எந்தெந்த சூழலில் பயன்படுத்தப் படுகின்றன என்பன போன்ற செய்திகளைத் தரும். சொற்பொருளடைவு (compendium) சொல்லாட்சி அகராதிகளினின்றும் ஒரு படி மேற்சென்று சொற்கள் வரும் இடங்களையும், சூழல்நிலைகளையும் பொருள் அடிப்படையில் பிரித்துத் தரும் கணிப்பொறியைப் பொருத்தவரையில் இரண்டையும் ஒன்றில் அடக்கித் தந்து விடலாம். புறநானூற்றிலிருந்து எடுக்கப்பட்ட ஒரு சொல்லின் சொற்பொருளடைவின் மாதிரி கீழேத் தரப்பட்டுள்ளன.

அணி (Sense)

1 அழகு

1.1 அழகு தாழ்ந்த அணி(புறம் 57.1)

1.2 அணிகொள் தேர் (புறம் 68.14)

.....

.....

.....

2.அணிகலன்

2.1 இழையணிப்பொலிந்த வேந்து கோட்டல் குல்(புறம் 81.1)

.....

.....

3. ஒப்பனை

3.1 ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண் அணியல் (பதி 51.9)

4. அணிமை

5. வரிசை

6. திரட்சி

7. ஆடை

8. ஒப்புமை

9. சூழ்தல்

10. பக்கரை (சேணம்)

சொல்லடைவுத் தயாரிக்க முன்னர் கூறியதுபோல் செய்யுளின் அடியில் சொற்களாகப் பிரித்துத் தரப்படவேண்டும். இத்துடன் சூழல் அடையாளப்படுத்தப்பட வேண்டும். வழியமைப்புகள் மூலம் நாம் சொல்லாட்சி அகராதியையோ சொற்பொருளடைவையோ தயாரிக்க இயலும். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்தி சங்ககால இலக்கியங்களுக்கு இத்தகைய சொல்லாட்சி அகராதியும் சொற்பொருளடைவும் தயாரித்துள்ளன.

2.13. தமிழ்ச் சொல்லாய்வி தயாரித்தல்

தமிழுக்கானச் சொல் ஆய்விகள் பல வெளிவந்துள்ளன. அவை தமிழ்ப்உரைகளை உள்ளீடு செய்து வடிவமைத்து வெளியீடாகப்பெறும் நுட்பங்களைக் கொண்டுள்ளன. word, word star போன்ற மென்பொருள்களில் கிடைக்கும் எல்லா வசதிகளும் இவற்றிலும் உள்ளன. சொற்களஞ்சியம், எழுத்து மற்றும் இலக்கணப் பிழைத்திருத்தி என்பன மேலும் மேம்படுத்தப்பட வேண்டும். பல தமிழ்ச்சொல் ஆய்விகளில் இவ்வசதிகள் இல்லை.

2.14. எழுத்துப் பெயர்ப்பு செய்தல்

ஒரு மொழியின் எழுத்துக்களால் எழுதப்பட்ட பிற மொழிகளின் எழுத்துக்களைப் பயன்படுத்தி படிக்கத்தக்க விதத்தில் மொழிப் பெயர்ப்புச் செய்யும் வசதிமுறைகள் கணிப்பொறியில் வந்து விட்டன. இதனால் தமிழ் எழுத்தில் எழுதப்பட்ட உரைகளை ரோமன் எழுத்தில் படிக்கவோ, ரோமன் எழுத்தில் எழுதப்பட்ட உரைகளை தமிழ் எழுத்தில் படிக்கவோ இயலும்.

2.15. தமிழுக்கு எழுத்துணரி உருவாக்குதல்

அச்சிடப்பட்ட உரைகளை அப்படியே கணினிக்கே எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியைப் பார்வை எழுத்தறிவான் (Optical charater recognitions) என்பர். இக்கருவி தட்டச்சு முறைபடி சாவிப்பலகை வழி உரைகளை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது. ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டு பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய

கருவியை உருவாக்குவது கடினமான செயல். இருப்பினும் தமிழுக்கு இத்தகைய பார்வை எழுத்துணரி உருவாக்கும் முயற்சிகள் உருவாக்கப்பட்டு ஓரளவுக்கு வெற்றியும் அடைந்துள்ளன. கிருண்ஷ்ண மூர்த்தி அவர்கள் தமிழுக்கு மிகத் திறமையான பார்வை எழுத்துணரி உருவாக்கியுள்ளார்.

2.16. தமிழில் கணிப்பொறி வழி அகராதி தயாரித்தல்

மொழி கற்றலில் கணிப்பொறியின் பங்கு மகத்தானது. ஒரு மொழியின் சொல் அதன் இலக்கணக் கூறு, பொருள், பயன்பாடு மற்றும் வாக்கியத்தில் அதன் உபயோகம் போன்றவைகளை அகராதி விளக்குகிறது. மேலும் ஒரு சொல்லுக்கு நிகரான பிறமொழி பதங்களையும் இருமொழி அகராதிகள் விளங்குகின்றன. அகராதி தயாரிக்கும் பணி எளிதானதல்ல. மிகுந்த நேரமும், அதிக உடல் உழைப்பும் பொருட்செல்வம் ஆகும். ஒரு அகராதியை உருவாக்க நீண்டகாலம் தேவைப்படும். எனவே அகராதி உருவாக்க கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

2.16.1. அகராதி உருவாக்கும் படிநிலைகள்

பொதுவாக அகராதி ஒன்றை தயாரிப்பதற்கு தேவைப்படும் படிநிலைகள்:

1. செய்திகளை சேகரித்தல், ஆராய்ந்து வகைப்படுத்துதல் அகர வரிசைப் படுத்துதல் என்பது முதற்கட்ட நிலைகளாகும்.
2. சொல்லுக்கான இலக்கணக்கூறு ஆராய்தல், பொருள் எழுதுதல், பிறமொழி சொற்களை எழுதுதல், சொல்லின் உபயோகத்தை விளக்குதல் இவைகளை இரண்டாம் நிலைப் பணியாகும்.

இதில் முதல் கட்ட பணி எளிமையாக கணிப்பொறி படுத்த இயலும். இரண்டாம் கட்டப்பணியை மனித முயற்ச்சியாலேயே செய்ய இயலும்.

2.16.2. கணிப்பொறி வழி அகராதி: வரலாற்றுச் சுருக்கம்

மாத்தூஸ் அர்னால் (Mathews Arnald) என்ற ஆங்கில கவிஞர் தன் கவிதைகளுக்கான சொல்லடைவை முதன் முதலாக கணிப்பொறி வழி உருவாக்கம் செய்தார். இதற்கு IBM-704 என்ற கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறி வழி சொல்லடைவினை முதன் முதலாக பேராசிரியர் ஸ்டிபன் (Stephen) உருவாக்கினார். இதனால் தொடர்ந்து Oxford அகராதி உருவாக்க சொல்லடைவு (Concordance) கணிப்பொறி வழி உருவாக்கப்பட்டது. இதைத் தொடர்ந்து ஆங்கிலம் மற்றும் பிற மொழிகளில் அகராதி உருவாக்க கணிப்பொறிகள் ஆங்காங்கே பயன்படுத்தப்பட்டன. தமிழ்ப்பல்கலைக்கழகத்தில் சங்க இலக்கியத்திற்கான' அகராதி உருவாக்க கணிப்பொறி பயன்படுத்தப்பட்டது. இதுவே முதல் முயற்சி ஆகும். கிரியா என்னும் தற்கால அகராதி கணிப்பொறி வழி உருவாக்கப்பட்டுள்ளது. இந்த அகராதி அச்சு வடிவிலும், குறுந்தகடுகளிலும் கிடைக்கப் பெறுகின்றன.

2.16.3. கணிப்பொறி வழி அகராதி: முறைமைகள்

கணிப்பொறி வழி அகராதி உருவாக்குவதற்கு கீழ்கண்ட முறைமைகள் பின்பற்றப்படுகின்றன.

1. செய்திகளை உள்ளீடு செய்தல்
2. செய்திகளை வார்த்தைகளாகப் பிரித்தல்

இந்நிலையில் ஒரு வாக்கியத்திலிருந்து சொற்கள் தனியாக பிரிக்கப்பட்டு பின்வரும் பட்டியல் உருவாக்கப்படுகிறது.

சொல்	வரி எண்
WORD	LINE NUMBER

1. மேற்கண்ட பட்டியலில் சொற்கள் அவைகள் இடம் பெறும் வரிசையில் இடம் பெறுகின்றன. மேலும் ஒரு சொல் உரையில் பல இடங்களில் பயனபடுத்தப்பட்டிருந்தால் அத்தனை முறை அச்சொல் இப்பட்டியலில் இடம்பெறும்.
2. மேற்கண்ட பட்டியல் அகரவரிசைப் படுத்தப்படுகிறது. இது மேற்கண்ட பட்டியலில் சொற்கள் இடம்பெறும் வரிசை முறையை மாற்றுகிறது. ஒரு சொல் பல முறை இடம் பெற்றிருப்பினும் அது அடுத்தடுத்து வரிசைமுறை பெறுகிறது.
3. இவ்வாறு அகரவரிசைப் படுத்தப் பட்டுள்ள பட்டுள்ள பட்டியலிலிருந்து கீழ்கண்ட பட்டியல் உருவாக்கப்படுகிறது.

சொல் (WORD)	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் இடம் LIST OF LINE NUMBER

இப்பட்டியலின் உதவிக் கொண்டு பின்வரும் பட்டியல் உருவாக்கப்படுகின்றது.

சொல் WORD	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் வரிசை எண் LIST OF LINE NUMBER	இடம் பெறும் வரி CITATION

மேற்கண்ட பட்டியலிலிருந்து ஒரு சொல் உரையில் எத்தனை முறை பயன்படுத்தப்பட்டுள்ளது என்பதை அறியலாம். அச்சொல் இடம்பெறும் வரி தெரிவதால் அதனைக்கொண்டு அச்சொல்லின் இலக்கணக்கூறு பொருள் உபயோகம் போன்றவைகளை ஆராய்ந்து எழுதலாம். இதன்மூலம் விவரங்கள் துல்லியமாகவும் விரைவாகவும் கிடைக்கப்பெறுகின்றன.

2.16.4. அகராதி தயாரித்தலில் இயற்கை மொழியாய்வு

கணிப்பொறி வழி அகராதி தயாரிக்கப்படும்போது கிடைக்கப்படும் இறுதி படிவம் சொல்லடைவு படிவம் போன்று இருக்கும். இதில் பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் தனித்தனி சொற்களாகவே இடம்பெறும். உதாரணமாக கணிப்பொறி, கணிப்பொறியை, கணிப்பொறியில், கணிப்பொறியால், கணிப்பொறிக்கு போன்ற சொற்கள் தனிச்சொற்களாகவே இடம்பெறும். இவைகளை சொல் வடிவங்கள் என மொழியியலாளர்கள் கூறுகின்றனர். ஆனால் அகராதியில் இவைகள் தனிச்சொற்களாக இடம்பெறும் (அல்லது) இடம்பெறாது. கணிப்பொறி என்ற பெயர்ச்சொல் மட்டுமே அகராதியில் இடம்பெறும். இது போன்ற அகராதி சொற்கள் மொழியியலாளர்கள் LEXMES என்கின்றனர். இது போன்று படி என்ற வினைச் சொல்லுக்கு படித்தான், படித்தார், படித்தது போன்ற எண்ணற்ற சொல் வடிவங்கள் இருப்பினும் படி என்பதே அகராதிச் சொல்லாகும். கணிப்பொறி வழி சொல்லடைவு தயாரிக்கும்போது சொல் வடிவங்களே இடம்பெறும். இதிலிருந்து அகராதிச் சொற்களை மனித முயற்சியினாலேயே மேற்கொள்ள வேண்டும். இப்பணியையும் முழுமையாக கணிப்பொறிப்படுத்த உருபனியல் ஆய்வு மேற்கொள்ள வேண்டும். இவ்வாறாக சொல் வடிவத்திலிருந்து (Word Forms) உருபனியல் பகுப்பாய்வு மூலம் அகராதி சொற்களை LEXEMS DICTIONARY WORDS உருவாக்கும் முறையை மொழியியலாளர்கள்

LEMMANTISATION என்று அழைக்கின்றனர். இம்முறையை கணிப்பொறிப்படுத்த முயற்சிகள் மேற்கொள்ளப்பட்டு வருகிறது. இதற்கு இயற்கை மொழி ஆய்வு உதவுகிறது.

2.16.5. முடிவுரை

கணிப்பொறி வழி அகராதி தயாரிக்கும் படிநிலைகளில் கணிப்பொறி படுத்த இயலும் நிலைகளையும், முறைகளையும் சுருக்கமாகப் பார்த்தோம். இம்முறைகளில் அகராதியில் மொழி ஒரு தடையில்லை. தமிழ் அகராதியோ, ஆங்கில அகராதியோ, பிறமொழி அகராதிகளையோ மேற்கண்ட முறைகளைப் பின்பற்றி உருவாக்கலாம். அந்தந்த மொழிகளில் அகர வரிசைப்படுத்தப் படும் மென்பொருள்களை வடிவமைக்க வேண்டும்.

2.17. தமிழில் கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல்

மக்கள் முன்னேற்றத்திற்கு கணிப்பொறி பல வகையலி உதவுகிறது. தொடக்கக் காலங்களில் கணிப்பொறி நல்ல கணக்கீட்டு கருவியாக (Calculating Device) பயன்பட்டது. அன்றாட பணிகளுக்கு அல்லது பணிகளிலும் வழக்கமாக செய்ய நேரிடும் பணிகளிலும் கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறியின் வேகம், துல்லியம், செய்திகளை மீளத்தரும் வேகம் மற்றும் நினைவகத் திறன் போன்றவை அதிகரித்தன. ஆனால் அதற்கு எதிர் மாறாக விலை குறைந்தது. அதன் விளைவாகக் கணிப்பொறி பயன்பாடு அதிகரித்தது.

ஆக்கப் பூர்வமான (creative) பணிகளுக்கும் கணிப்பொறிகள் பயன்படும் வண்ணம் உருவாக்கப்படுகின்றன. குறிப்பாகக் கல்விக் கூடங்களில் பயிற்றுவிப்பதற்கு உதவும் கருவியாக கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன. கணிப்பொறி வழி

பாடங்களை தயாரித்து உரிய திட்டங்களை செயல்படுத்தி மாணவர்களுக்கு கற்பிக்கலாம். பாடங்களைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறி உதவுகிறது.

2.17.1. கற்றல், கற்பித்தலில் கணிப்பொறி

அடிப்படையில் கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. தேவையான பணிக்கான வழியமைப்பு அல்லது மென்பொருள் கொண்டு அது இயக்கப்படுகிறது. இருப்பினும் செய்திகளை சேகரித்தல், ஆராய்தல், வெளியிடுதல் மற்றும் பரப்புதல் போன்ற செயல்பாடுகளில் மிக மிக அதிவேகமுடையது. மேலும் கணிப்பொறியில் உரைச் செய்திகள், வரைபடங்கள் புகைப்படங்கள் ஒலிச்செய்திகள் மற்றும் சலனப்படங்கள் (animation) போன்ற பலவகையான செய்திகளை உள்ளீடு செய்து ஆராய்ந்து முடிவு பெறுவதற்கான வசதிகள் உள்ளன. எல்லா வகை செய்திகளையும் கலந்த கலவையாகவும் செய்திகள் அமையப்பெறலாம். இது போன்று எல்லா வகை செய்திகளையும கலந்த கலவையை பல்லூடகம் (multimedia) என்கிறோம்.

மற்றொரு பக்கத்தில் பயிற்றுவிப்போர் அல்லது ஆசிரியர்களுக்கு இருக்க வேண்டிய அடிப்படை இயல்புகள் மூன்று அவையாவன:

1. செய்திகளைத் திரட்டுதல்
2. செய்திகளை ஆராய்தல்
3. பயில்வோருக்கு ஏற்றவாறு செய்திகளை எடுத்துரைத்தல்

இத்தகைய குணங்களைக் கணிப்பொறி பெற்றிருப்பதால் கற்பித்தலுக்கும் கற்பதற்கும் சிறந்த சாதனமாகப் பயன்படுத்தப்படுகிறது. கணிப்பொறியால் சொந்தமாகப் பாடங்களை உருவாக்கவோ பாடங்களை நல்ல முறையில்

வடிவமைப்பதற்கோ இயலாது. இவ்வாறான கணிப்பொறியை ஒரு ஆசிரியரைப் போன்று செயல்பட வைப்பதற்கு மென்பொருள் (soft ware) உருவாக்கப்படுகின்றன. ஒரு பாடத்தைப் பயிற்றுவிப்பதற்காக உருவாக்கப்படும் மென்பொருள் பொதுவாக பாட மென்பொருள் (course ware) எனப்படுகிறது.

கணிப்பொறி வழி கற்பித்தலுக்கு CAI (Computer Assisted/ Aided Instruction) என்று பெயர். இங்கு Assisted (or) Aided என்ற சொல் கற்பித்தலுக்கு கணிப்பொறி உதவும் என்ற கருத்தை வலியுறுத்தும். அதுபோல் கணிப்பொறி மூலம் பயிலும் முறைக்கு CAL (Computer Aided Learning) என்றுப் பெயர். CAL and CAI போன்ற சொற்களுக்கு நிகராக பல தொடர்கள் பயன்படுத்தப் படுகின்றன. CMI (Computer Managed Instruction), CML (Computer Managed Learning), CDL (Computer Directed Learning) என்பன சிலவாகும்.

2.17.2. கணிப்பொறி வழி கற்றல்-கற்பித்தலின் சிறப்பியல்புகள்

கணிப்பொறி வழி கற்றல், கற்பித்தலில் எண்ணற்ற பயன்களும், சிறப்பியல்புகளும் உள்ளன. பொதுவாக இச்சிறப்பியல்புகளை மூன்று வகைகளாகப் பிரிக்கலாம். அவையாவன,

1. கணிப்பொறியில் இயல்பாக இருக்கின்ற சிறப்பு குணங்கள்.
2. பயில்வோருக்கு கிடைக்கின்ற பயன்கள்.
3. பயிற்றுவிப்போருக்கு உதவும் சிறப்புகள்.

கணிப்பொறியின் செயல், வேகம், துல்லியம், சோர்வின்மை, சகிப்புத்தன்மை, நெகிழ்வுத் தன்மை, ஒரே நேரத்தில் பல செயல்களை புரியம் திறமை போன்றவை

கணிப்பொறியில் இயல்பாக அமைந்துள்ள பண்பு நலன்களாகும். கணிப்பொறிகள் மனிதனைப் போல் சோர்வடைவதில்லை. எனவே எப்பொழுது வேண்டுமானாலும் எவ்வளவு நேரம் வேண்டுமென்றாலும் தொடர்நது பயன்படுத்தலாம். எனவே, கணிப்பொறிகள் கல்வி கற்றல் கற்பித்தலில் பயன்படுத்தப் பட்டு வருகின்றன. மேலும் வேறு எந்த ஒரு கருவியிலும் இல்லாத பல்லூடக சாதன அமைப்புக் கொண்டது. எனவே உரைநடைச் செய்திகள் மட்டுமல்லாது வரைப்படங்கள், புகைப்படங்கள் மற்றும் ஒலிச் செய்திகள் போன்றவைகளை ஆராய்வதற்கும், கையாளவதற்கும் எண்ணற்ற வசதிகளைப் பெற்றுள்ளது. இதனால் கல்வி கற்பிப்பதற்கும் கற்பதற்கும் தேவையான அனைத்து அடிப்படை வசதிகளையும் பெற்றுள்ளது.

பட வில்லைகள் (Slides) புகைப்படங்கள் மற்றம் கல்வித் தொடர்பான சினிமா படங்கள் கற்றலுக்கும், கற்பித்தலுக்கும் சிறந்த சாதனங்களாக பயன்படுத்தப்பட்டு வருகின்றன. இவைகளை கருவியாக பயன்படுத்திக் கொள்ள இயலுமெனினும் அந்த நிகழ்ச்சிகளுடன் உரையாடுவதற்கோ, கற்போரின் தன்மைக்கேற்ப கட்டுப்படுத்தவோ இயலாது. அந்நிகழ்ச்சிகளை தயாரித்த வரிசை கிரம முறையில் மட்டுமே பார்க்கலாம். கற்போரின் மனநிலைக்கு ஏற்ப தேவையான செய்திகளை கூடுதலாகவோ அல்லது குறைவாக பெறவோ இயலாது. ஆனால் கணிப்பொறி வழி இத்தகைய கல்வி நிகழ்ச்சிகள் அல்லது பாடங்களைப் பதிவு செய்யும் பொழுது அதனுடன் உரையாட இயலும். தேவையான செய்திகளை தேவையான வண்ணம் கேட்டுப் பெறலாம். இது கணிப்பொறி கருவியின் மிகப் பெரிய சிறப்பம்சமாகும்.

கணிப்பொறியை எந்நேரமும் எப்பொழுதும் பயன்படுத்தலாம் என்பதால் பயில்வோர் விருப்பமான நேரத்தில் விருப்பமான பாடத்தைப் பயிலலாம். கணிப்பொறி வழி பயில்வோரின் சுரந்திரம் தேவை. இங்கு பயில்வோர் தனித் தனியாக பயில வாய்ப்புகள் உண்டு. 1.பயில்வோரின் ஈடுபாடு, 2. கற்கும் திறன், 3. கற்கும்

வேகம் ஆகிய தன்மைகளைப் பொறுத்து பாடங்களைக் கற்றுக் கொள்ளலாம். சாதாரண வகுப்புகளில் இத்தகைய சாத்தியங்கள் இல்லை. பாடங்களை விரைவாக கற்றுக் கெர்ளபவர்கள் மெதுவாக கற்கு கொள்பவரின் வேகத்திற்கு ஒத்துப்போக வேண்டும். இதனால் வேகமாக கற்றுக் கொள்பவரின் நேரம் வீணாகி போகும். வகுப்பில் கூச்ச சுபாவத்தால் ஐயப்பாடுகளை களைவதற்கான கேள்விகளை கேட்பதற்கு மாணவர்கள் தயங்குவர். ஆனால் கணிப்பொறி வழி கற்கின்ற பொழுது மனநிலையைப் பொருத்து ஒரு பாடத்தில் அதிக செய்தியையோ, பொருளையோ, முக்கிய செய்தியையோ படித்து அறிவதற்கான பரிமாணங்களை அல்லது சிறப்புகளைப் பயில்வோரே நிர்மானித்துக் கொள்ளலாம். மேற்கூறியவைகள் எல்லாம் கணிப்பொறி வழி பயில்வோருக்கு உள்ள சிறப்புத் தன்மைகள் ஆகும்.

கணிப்பொறி வழி கற்பிப்பதற்கான பாடங்களை உருவாக்கும் ஆசிரியர்கள் மாணவர்களின் கல்வி நிலைக்கு ஏற்ப, கற்கும் திறனுக்கு ஏற்ப வடிவமைக்க வேண்டடும் என்பதில்லை. ஒரு பொதுவான பாடங்களைத் தயாரித்து பயில்வோரின் திறமைக்கு ஏற்ப அவைகளை உருவாக்கித் தர இயலும். உரைநடைச் செய்திகள், விளக்கப் படங்கள் அனைத்தும் இதில் உள்ளதால் பாடக் கருத்துகளை எளிதாக புரிய வைக்க இயலும். ஒருமுறை பாடத் திட்டங்களை வகுத்து செம்மையாகத் தயாரித்து கணிப்பொறியில் எத்தனை முறை வேண்டுமானாலும் பயன்படுத்தலாம். ஆயிரக்கணக்கானவர்க்கு இத்தகைய பாடங்கள் பயன்படுத்துவதற்கு அமையும். இத்தகைய பாடங்களை ஒரு துறை நிபுணத்துவம் பெற்ற ஆசிரியர்கள் உருவாக்குகின்றனர். அவைகளில் திறமையானது ஆயிரக் கணக்கானோருக்கு உலகம் முழுவதும் பயன்படும்.

2.17.3. CAI- உருவாக்குவதற்கான கணிப்பொறி மொழிகள்

CAI – கணிப்பொறி வழி கற்பித்தல் மற்றும் கற்றல் பயன்பாடுகளுக்கான மென்பொருள்களை உருவாக்குவதற்கு என்றே கணிப்பொறி அமைப்பு மொழிகள் உள்ளன. IBM – (Internet Bussiness Machine) என்ற நிறுவனம் Course Writer என்ற வழியமைப்பு மொழியை இதுபோல் Control Data Corporation என்ற நிறுவனமும் TW என்ற மொழியை Starkwealther என்ற நிறுவனமும் Extol என்ற மொழியை Kenning and Kenning என்ற நிறுவனமும் தயாரித்து வெளியிட்டுள்ளன.

கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான பாடங்களை உருவாக்க மேற்கூரிய கணிப்பொறி மொழிகள் உதவுகின்றன. இவைகள் ஏனைய கணிப்பொறி மொழிகளை விட எளிமையானவை. இருப்பினும் ஒரு கணிப்பொறி மொழி கற்பதற்கான அதை பயன்படுத்துபவர்களுக்கான திறன் பெற்றவர்க்ள மட்டுமே இவைகளை பயன்படுத்த இயலும். எனவே இவைகளைக் கொண்டு பாடங்களை உருவாக்குபவர்கள் இக்கணிப்பொறி மொழியில் தேர்ச்சி பெறுவது அவசியம் ஆகும். மேலும் வழியமைப்பு திறன் கொண்டவர்களாகவும் இருத்தல் அவசியம்.

2.17.4. CAI உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள் (packeges)

பொதுவாக வழியமைப்பு மொழிகளை பயன்படுத்தி ஒரு திட்டங்களை உருவாக்குகின்றபொழுது திட்டங்களை செயல்படுத்துவதற்கான முறைமைகளில் தேர்ச்சி பெற்றிருக்க வேண்டும். உதாரணமாக பத்து சொற்களை கணிப்பொறிக்கு தந்து அகரவரிசை பயன்படுத்துவதற்காக கொள்வோம். இப்பணியை செய்ய வழியமைப்பு மொழியில் தேர்ச்சி மட்டும் போதாது. அகரவரிசைப் படுத்துவதற்கான முறைகளை தெரிந்தவராகவும் இருத்தல் வேண்டும். அவ்வாறு இருந்தால் தான் அந்த முறைகளையும், செயல்பாடுகளையும் கணிப்பொறிக்கு கட்டளைகளாக செயல் படுத்த இயலும். ஆனால் பயனீட்டு மென்பொருள்களில் இதுபோன்ற செயல்பாடுகளுக்கு

தேவையானவை தயார் நிலையில் உள்ளன. அவைகளைப் பயன்படுத்தி கொள்வதற்கான சாதாரண திறமைகள் இருந்தால் போதுமானது. எனவே, கணிப்பொறி வழி கற்றல் கற்பிப்பதற்கான மென்பொருள்கள் உருவாக்க கணிப்பொறி மொழிகளை பயன்படுத்த முனைவோர் பாடத் திட்டங்களை மட்டுமல்லாது கணிப்பொறி மொழியிலும் தேர்ச்சிப் பெற்றவராக இருத்தல் வேண்டும். இது ஒரு மிகப்பெரிய தடையாகும். இத்தடைகளை அகற்றவும், இப்பணிகளை எளிமைப் படுத்தவும் பல பயனீட்டு மென்பொருள் (Utility Soft Ware) வெளிவருகின்றன. இவைகளில் பாடங்களை கணிப்பொறி வழி தயார்படுதற்கான அனைத்து வசதிகளும் தயார் நிலையில் உள்ளன. இத்தகைய பயனீட்டு மென்பொருள்களை எப்படி பயன்படுத்திக் கொள்வது என்ற பயிற்சி மட்டுமே போதுமானது. இவைகளை எளிதாகக் கற்றுக் கொள்ளலாம். கணிப்பொறி பயிற்சி இல்லாதவாகள் எளிமையாகவும், விரைவாகவும் கற்றுக் கொள்ள வசதி உள்ளது. Story Board மற்றும் Copy Write போன்றன CAI உருவாக்குவதற்கு பயன்படுத்தப்படும் பயனீட்டு மென்பொருள்கள் ஆகும். இவைகள் தவிர Light Source என்ற மென்பொருள்களும் உள்ளது.

2.17.5. CAI–இல் நகரும் படங்கள் (animation)

CAI மென்பொருள் உருவாக்கத்தில் நகரும் படங்கள் அல்லது சலனப்படங்கள் மற்றும் வீடியோ காட்சிகள் போன்றவைகளை அமைப்பது மென்பொருளின் தரத்தை மேம்படுத்தும். இதற்காக சில மென்பொருள்கள் உள்ளன. அவற்றுள் VIDEO WORKS என்பது ஒன்றாகும். இம்மென்பொருளில் சலனப்படங்களை உருவாக்கவும் கையாளவும் தேவையான வசதிகள் உள்ளன.

2.17.6. CAI உருவாக்கத்தில் மனித நுண்ணறிவின் பங்கு

கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. மனிதனைப் போன்று சிந்தித்து செயல்படக் கூடிய கணிப்பொறிகளை உருவாக்க அறிஞர்கள் முனைற்துள்ளனர். கணிப்பொறியை மனிதரைப் போல் உருவாக்கவும், சூழலுக்கு ஏற்ப பணியாற்றவும் அனுபவத்தில் புதிய செய்திகளை சுயமாக கற்கவும் தேவையான திறமைகளை உருவாக்கி செயல்படுமாறு உருவாக்குகின்றன. இத்துறைக்கு செயற்கை நுண்ணறிவு (artificial intelligence) என்று பெயர். இதை AI என்று சுருக்கமாக அழைப்பர். இத்துறையின் ஒரு அங்கமாக இயற்கை மொழியாய்வு NLP (Natural Language Processing) என்ற துறை உள்ளது. ஒரு மனிதன் மற்றவர்களிடம் உரையாடுவது போல் மனித மொழியில் கணிப்பொறியுடன் உரையாட தேவையான வசதிகளை இத்துறை செய்கிறது. இந்த ஆய்வு வெற்றி பெறுமேயானால் கணிப்பொறிக்கு கட்டளையிட கணிப்பொறி மொழி என்ற ஒன்று தேவையில்லை. மனித மொழியில் கணிப்பொறிக்கு கட்டளை தரலாம். அதை கணிப்பொறி புரிந்துக் கொண்டு செயலாற்றும். அதுபோல் கணிப்பொறியின் விடையை மனித மொழியில் விளக்கமாகத் தரும். இத்தகைய அமைப்புகள் மனிதனுக்கும், கணிப்பொறிக்கும் இருக்கின்ற உறவை எளிமைப்படுத்தும். மேலும் துறை ஒருங்கு (Expert System) ES என்பது கணிப்பொறி எழுதியுள்ள மிகப்பெரிய பிரிவாகும். கணிப்பொறியை ஒரு துறையின் வல்லுநரைப் போன்று பயன்படுத்த இவ்வாய்வுகள் உதவும்.

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருள்களை உருவாக்கும் பொழுது மேற்கூறிய மொழி நுட்பங்கள் மனித நுண்ணறிவு இயற்கை மொழி ஆய்வு துறை ஒருங்கு போன்றன சேர்க்கப்படுகின்றன. செயற்கை நுண்ணறிவுடன் கூடிய CAI மென்பொருளை ACAI என்று அழைக்கிறார்கள். இதை ITS (Intelligence Tutorial System) என்றும் அழைப்பர்.

2.17.7.கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் கணிப்பொறி சிறந்து பங்காற்றுகிறது. ஆனால் மொழி கற்றல் கற்பித்தல் போன்ற செயல்களில் எண்ணற்ற நுணுக்கங்கள் தேவைப்படுகின்றன. இதில் கணிப்பொறி ஓரளவுதான் பங்காற்ற முடியுமே தவிர பயிற்று விக்கும் ஆசிரியருக்கு நிகராக அமையும் அல்லது அமையாது. மொழி கற்பிப்பதற்கு கணிப்பொறி துணை கருவியாக மட்டும் பயன்படும். இருப்பினும் இத்தகைய மென்பொருள் உருவாக்குகின்ற பொழுது எது இயலுமோ அதில் உறுதியும், திடமான சிந்தனையும் வேண்டும். மொழி கற்றல் கற்பித்தலில் கணிப்பொறியில் செய்யத் தரும் பயிற்சிகள் கீழேப் பட்டியலிடப் பட்டுள்ளன:

- கணிப்பொறி வழி உரையாடல் பயிற்சி அளித்தல்.
- மொழியில் கோடியிட்ட இடங்களை நிரப்புதல் போன்ற வினாத்தாள்களின் பயிற்சி.
- ஒரு சொல்லைத் தந்து அதற்கான உச்சரிப்பு.
- வாக்கிய வகைளை மாற்றி அமைத்தல்.
- இரு சிறு வாக்கியங்களை இணைத்து ஒரு வாக்கியமாக உருவாக்குதல்.
- சொல்லின் ஒருமை, பன்மை, எதிர்பதங்கள் போன்றவற்றை கற்று கொள்வதற்கான பயிற்சிகள்.
- சொற்களை பிரித்து எழுதுதல்
- சொற்களை சேர்த்து எழுதுதல்
- ஒரு சொல்லின் இலக்கணக் கூறினைக் கொடுத்து மற்ற இலக்கணக் கூறுகளைக் கேட்டல்.

மேற்கூறிய பட்டியலில் கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான சில வாய்ப்புகள் கொடுக்கப்பட்டுள்ளன. இதுபோன்று எளிமையான பயிற்சிகளை தருவதற்கு மென்பொருள்களை உருவாக்கி பயன்படுத்தலாம்.

2.17.8.கணிப்பொறி வழி தமிழ் கற்றல்-கற்பித்தலுக்கான மென்பொருள்கள்

கணிப்பொறி வழி கறிபத்தலுக்கு எண்ணற்ற மென்பொருள்கள் உள்ளன. அறிவியல் தொழில் நுட்பம் போன்ற பாடங்களை பயிற்றுவிப்பதற்கு உலகளவில் பல்லாயிரக்க கணக்கான மென்பொருள்கள் உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன. அதுபோல் மொழி கற்றல் கற்பித்தலுக்குப் பல மென்பொருள்கள் உள்ளன. கணிப்பொறி வழி தமிழ் கற்றல் கற்பித்தலுக்கு மெனபொருள் உருவாக்கம் தொடங்கப்பட்டுள்ளது.

சென்னையைச் சார்ந்த Learn Fun என்ற நிறுவனம் தமிழுக்காக மென்பொருள் ஒன்றை உருவாக்கி உள்ளது. இம்மென்பொருளில் தமிழ் எழுத்துக்களை எழுத, பழக வசதிகள் உள்ளன. இம்மென்பொருள் மொழி விளையாட்டு மென்பொருள் வகையைச் சேர்ந்தது. தமிழை ஆரம்ப நிலையில் கற்பவர்கள் வல்லினம், மெல்லினம், குறில், நெடில் போன்ற அடிப்படைக் கூறுகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் மிகவும் உதவுகிறது.

சென்னையைச் சார்ந்த Soft View என்ற நிறுவனம் அமுதம் என்ற மெனபொருளை உருவாக்கியுள்ளது. இம்மென்பொருள், திருக்குறள், ஆத்திச் சூடி போன்றவைகளை எளிமையாகக் கற்று தருகிறது. சிங்கப்பூர் தேசிய கல்விக் கழகம் தமிழில் பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை உருவாக்கியுள்ளது. ஒருமை, பன்மை, எதிர்ப்பதம் போன்றவைகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் உதவுகிறது.

சிங்கப்பூரைச் சார்ந்த தமிழாசிரியர் கழகம் தமிழைப் பயிற்றுவிப்பதற்காக சில மென்பொருளை வெளியிட்டுள்ளன. அவைகளைக் கொண்டு தமிழில் உள்ள இரண்டு

எழுத்துச் சொற்கள், மூன்று எழுத்துச் சொற்கள் போன்றவைகளை எளிமையாக கற்கலாம். மேலும் வெளிப்படும் படத்திலிருந்து அப்படத்தில் உள்ள பொருளை அடையாளம் காண்டு பயிற்சிப் பெற சில வசதிகள் இம்மென்பொருளில் உள்ளன. மேலும் ஒன்றியிலிருந்து ஆயிரம் வரை எண்கள் எழுதவும், படிக்கவும் சொல்லித் தருகிறது. மேற்கண்ட மென்பொருள் தயாரிக்கும் பணி அமரர் பேராசிரியார் கோவிந்தசாமி தலைமையில் சிங்கப்பூரில் நடைப்பெற்றது.

அமெரிக்காவைச் சார்ந்த லிண்ட்ஹோம் என்பவர் தமிழ்ச் சொற்கோவைகளை பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை வெளியிட்டுள்ளார். பென்சில்லேனியா பல்கலைக் கழகத்தில் பேச்சுத் தமிழைப் பயிற்றுவிப்பதற்காக மென்பொருள் உருவாக்க முனைந்துள்ளனர். அமெரிக்காவில் உள்ள சிக்காக்கோ நகரிலுள்ள நிறுவனம் ஒன்று Visual Tamil என்ற மென்பொருளை உருவாக்கியுள்ளது. இம்மென்பொருள் தமிழ் கலை, பண்பாடு ஆகியவற்றைக் கற்பிக்கிறது. இது போன்று ஆங்காங்கே கணிப்பொறி வழி தமிழில் கற்பித்தல் கற்றலுக்கு பல மென்பொருள்கள் உருவாக்கப்படுகின்றன.

2.17.9. கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு

ஒன்றுக்கு மேற்பட்ட கணிப்பொறிகள் ஒன்றோடு ஒன்று இணைக்கப்படுகின்றன. இத்தகைய அமைப்பிற்கு கணிப்பொறி வலைப்பின்னல் (Computer Net work) என்று பெயர். இதனால் ஒரு கணிப்பொறியில் பதிவு செய்யப் பட்டுள்ள மென்பொருள் மற்றும் பதிவுகளை கணிப்பொறிகளில் பெறலாம். இது போன்று உலகளாவிய நிலையில் உள்ள கணிப்பொறிகளை ஒன்றிணைத்து தகவல் பரிமாற்றம் செய்து கொள்ள உதவும் அமைப்பிற்கு இணையம் என்று பெயர். இதன் பயனாய் இணையத்தில் உள்ள ஒரு கணிப்பொறியில் பதிவு செய்யப்படும் செய்திகளை உலகில் உள்ள எந்த கணிப்பொறியிலிருந்தும் பெறலாம். கணிப்பொறி வழி மொழி கற்றல்

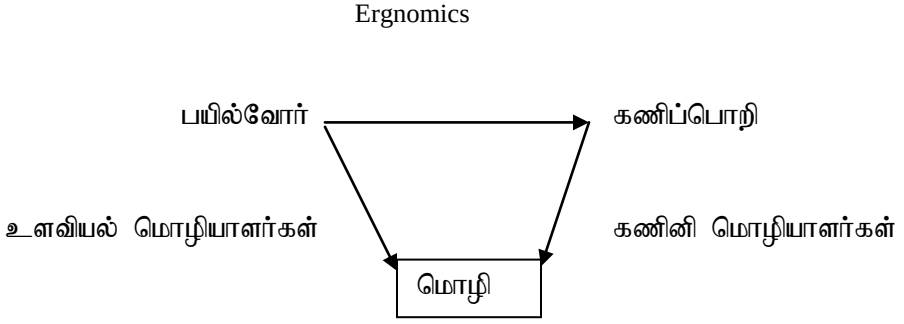
கற்பித்தலுக்கான மென்பொருள் இணையம் ஏற்றம் செய்யப்பட்டுள்ளன. இம்மென்பொருள்களை செயல்படுத்தி மொழிகளை கற்பிக்கலாம். பென்சில்லேனியா பல்கலைக் கழகத்தில் WALTT – Web Aided Language Teaching and Learning அல்லது Training என்ற திட்டம் செயல்படுத்தப் பட்டுள்ளது.

இணையத்தின் வழி பேச்சு தமிழை பயிற்றுவிப்பதற்கு இது உதவுகிறது. மேலும் தமிழ் மொழியை கற்பிப்பதற்கு சிங்கப்பூரில் உருவாக்கப்பட்டுள்ள பல இணையத் தளங்கள் செம்மையாக உள்ளன. தமிழ்ப்பல்கலைக் கழகத்தின் அங்கமாக தமிழ் இணையப் பல்கலைக்கழகம் செயல்படுகிறது. இணையப் பல்கலைக்கழகம் தமிழ் பாடத்தில் சான்றிதழ் வகுப்புகளையும், பட்டய வகுப்புகளையும், தொலைத்தூர கல்வி திட்டத்தின் கீழ் இணையம் வழி நடத்துகிறது. அயல் நாட்டில் உள்ள தமிழர்கள் மற்றும் இந்திய மாநிலங்களில் உள்ள தமிழர்கள் தமிழ் மொழியை இதன் வழி கற்கலாம். இப்பல்கலைக் கழகம் தமிழ் மொழி கலாச்சாரம், பண்பாடு போன்றவற்றை இணையம் வழி கற்பிக்கிறது. மேலும் தமிழ் மொழியை கற்றுக் கொள்வதற்கு உறுதுணையாக அமையும் அகராதிகள், இலக்கியங்கள் போன்றவைகளை இந்த இணையத்தின் மூலம் இலவசமாக கற்கலாம். இது போன்று தமிழ்மொழி கற்றலுக்கும், கற்பித்தலுக்கும் இணையத்தில் பல வாய்ப்புகள் உள்ளன.

2.17.10. மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு

இத்தகைய மென்பொருளை உருவாக்க மொழி வல்லுநர் உளவியல் மொழியியாளர்கள் (Psychology Linguistics) கணினி மொழியியல் வல்லுநர்கள் (Computational Linguistics) ஆகியோரின் ஒத்துழைப்பு அவசியமாகும். மொழியியல் வல்லுநர்கள் கணிப்பொறிக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். உளவியல் மொழியாளர்கள் பயில்வோருக்கும், மொழிக்கும் உள்ள உறவை

ஆராய்கின்றனர். மேலும் கணிப்பொறிக்கும், பயில்வோருக்கும் உள்ள உறவை ERGNOMICS என்ற துறையினர் ஆராய்கின்றனர். இத்தொடர்பு கீழே கண்ட கட்டப்படம் விளக்குகிறது.



கணிப்பொறி வழி கற்றல், கற்பித்தலில் மேற்கூறியவாறு கணிப்பொறி பல வகைகளில் பயன்படுகிறது.

2.18. கணிப்பொறி வழி மொழிபெயர்ப்பு

பாரதத்தில் பன்மொழி பேசும் மக்கள் உள்ளன. மொழி, இனம், பண்பாடு மற்றும் கலாச்சாரத்தால் பல்வேறு மக்களின் எண்ணங்கள், கருத்துக்கள் பிறமொழி பேசும் மக்களிடையே சென்றடைவது அவசியமாகும். அதுபோல் ஒருமொழி பேசும் மக்களின் பண்பாடு, கலாச்சாரம், இலக்கியம் போன்றவைகளாகும். பிறமொழி பேசும் மக்களிடையே சென்றடைய வேண்டும். அப்பொழுதுதான் மக்களிடையே ஒற்றுமை உணர்ச்சியும் தேசிய ஒருமைப்பாடும் நிகழும். ஒரு மொழியில் உள்ள செய்திகள் பிற மொழிகளில் மாற்றம் செய்யப்படுவது மிகவும் அவசியமாகும். இதுபோல் உலக நாடுகளில் ஒரு நாடு பிற நாடுகளிலிருந்து தனித்து இருக்க இயலாது. தொழில் நுட்ப மேம்பாடு, பொருளாதார மேம்பாடு, சமூக மேம்பாடு போன்றவைகளுக்கு ஒன்றோடு ஒன்று தொடர்பு கொண்டு கூட்டு முயற்சியில் மனித வாழ்வியலில் மேம்பாட்டினைக் கருத்தில் கொள்ள வேண்டும்.

ஒரு நாட்டின் கலாச்சாரம் பண்பாடு போன்ற பிற நாடுகளுக்கும் பரவ செய்ய வேண்டும். ஒரு நாட்டின் இலக்கியச் செல்வங்கள் பிற நாடுகளைச் சென்றடைய வேண்டும். அப்பொழுதுதான் ஒரு நாட்டின் தொன்மை, தனித்துவம், வளம் போன்றவைகள் பிற நாடுகளுக்கு தெரிய வரும். எனவே ஒரு நாட்டில் உருவாக்கப்படும் படைப்புகள் பிற நாடுகளுக்கு எளிதில் சென்றடைய வேண்டுமெனில் மொழிபெயர்ப்புப் பணி அவசியமாகும். நாம் தகவல் யுகத்தில் வாழ்ந்து வருகிறோம். செய்திகள் மலையெனக் குவிகின்றன. இச்செய்திகளை உடனுக்குடன் மொழிபெயர்க்க வேண்டியது அவசியமாகிறது. வாழ்வில் எல்லா நிலைகளலும் பயன் தரக் கூடிய கணிப்பொறி மொழி பெயர்ப்புக்குப் பயன்படுத்த மேற்கொள்ளப் பட்டுள்ளன.

ஒரு மொழியில் உள்ள செய்திகளை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பு இயல்பு தன்மை மாறாது இருத்தல் வேண்டும். செய்தி எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Languages) என்கிறோம். செய்திகள் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியை இலக்கு மொழி என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியில் மேற்கொள்ளுவதென்பது மிகக் கடினமானப் பணியாகும். எனவே மொழி பெயர்ப்புகளை கணிப்பொறி வழி மொழிபெயர்ப்பு (Machine Translation) மேற்கொள்வதற்கான தேவைகள் மற்றும் அவசியங்கள் கீழே பட்டடியலிடப் பட்டுள்ளன:

- உலக அளவில் செய்திகள் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழி பெயர்த்தல் அவசியமாகிறது.
- கணிப்பொறி வழி மொழி பெயர்த்தல் மனித உழைப்பை குறைக்கிறது.

- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்பு பணிகளை செய்ய இயல்கிறது. இதனால் கால விரயம் தவிர்க்கப் படுகிறது.
- தற்போதைய மொழி பெயர்ப்பு பணிகளுக்கு ஏற்ற வகையில் மொழிபெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மூலமொழி பெயர்க்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழி மற்றும் இலக்கு மொழிகளில் புலமை வாய்ந்தவராக இருத்தல் வேண்டும். ஆனால் கணிப்பொறியில் இந்தக் கட்டுபாடு இல்லை. தேவையான மொழியில் புலமை உள்ளவர்களிடம் கருத்துகள் பெற்று அதை வழியமைப்பு மொழிகளில் எழுதி மொழி பெயர்ப்பு செய்யலாம்.

2.18.1. கணிப்பொறி மொழி பெயர்ப்பின் வகைகள்

பொதுவாக கணிப்பொறி வழி மேற்கொள்ளப்படும் மொழிபெயர்ப்பு பின் வரும் மூன்று வகைகளாகப் பிரிக்கப்படுகின்றன:

1. கணிப்பொறி உதவியுடன் கூடிய மனித மொழி பெயர்ப்பு (CAHT – Computer Assisted Human Translation)
2. மனித உதவியுடன் கூடிய இயந்திர மொழி பெயர்ப்பு (HAMT – Human Assisted Machine Translation)
3. முற்றிலும் தானியங்கி (அல்லது) இயந்திர மொழிபெயர்ப்பு (AMTS – Automatic Machine Translation System)

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனித மொழி பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகிறது. அவர்கள் மொழி பெயர்ப்பு பணியில் ஈடுபடும் போது ஒரு மொழியில் உள்ள சொற்களுக்கு நிகரான பிறமொழிச் சொற்களை தெரிந்து கொள்ளவும் அச்சொற்களின் பொருள், பயன்பாடு ஆகியவற்றை எளிமையாக தெரிந்து

கொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாக தயார் நிலையில் தெரிந்து கொள்வதற்கு கணிப்பொறி உதவுகிறது. இவ்வகை மொழிப்பெயர்ப்புகளில் மொழிப்பெயர்ப்பாளர்கள் முதல் நிலையிலும், அவர்களுக்கு உதவக் கூடிய வகையில் கணிப்பொறி பயன்படுகிறது.

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிபெயர்ப்பு பணியை செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ளன. தகவலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிபெயர்ப்பு வகைகளில் மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பே நடைமுறையில் உள்ளது.

முற்றிலும் இயந்திர மொழி பெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல. இதற்கு எண்ணற்ற மொழியின் கோட்பாடுகளும், மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட வேண்டியது கற்று இருப்பினும் கணிப்பொறி வழி மொழி பெயர்ப்பில் முற்றிலும் தாணியங்கப் படுத்தும் முயற்சி இலக்காக கொள்ளப்பட்டுள்ளது.

2.18.2. கணிப்பொறி வழி மொழி பெயர்ப்பின் வரலாறு

இயந்திர மொழிபெயர்ப்பு அமைப்பொழுங்கின் வளர்ச்சி

இயந்திர மொழிபெயர்ப்பு அமைப்பொழுங்கின் வளர்ச்சி கடந்த நூற்றாண்டுகளில் செய்யப்பட்ட ஆராய்ச்சி முன்னேற்றம் குறித்தக் காரணிகளால் ஊக்கப்படுத்தப்பட்டுள்ளது. இயந்திர மொழிபெயர்ப்பின் உண்மையான பிறப்பு

இரண்டாவது உலகப்போருக்குப்பின் டிஜிட்டல் கணிப்பொறியின் வரவால் விளைந்தது. இயந்திர மொழிபெயாப்பின் வளர்ச்சியை நான்கு கட்டமாகப் பகுத்து ஆயலாம். (Hutchins 1986).

முதல் கட்டம் (1946 to 1982)

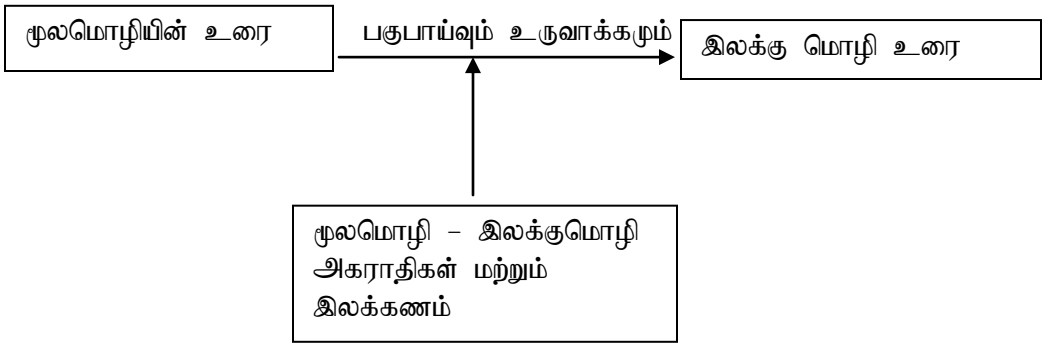
இந்தக் காலகட்டதக்தில் செய்யப்பட்ட இயந்திர மொழிபெயர்ப்புகள் மூலமொழி சொற்களுக்கு இணையானச் சொற்களை இருமொழி அகராதியில் பார்த்து கண்டுபிடித்து மூலமொழிச் சொற்களுக்கு இணையான சொற்களை மூலமொழி போலவே வரிசைப்படுத்தித் தருவதைச் செயல் முறையாகக் கொண்டிருந்தது. இச்செயல் முறை தொடரியல் பகுத்துக்குறித்தலைப் பயன்படுத்தவில்லை.

இரண்டாவது காலகட்டம் (1954-1966)

1954இல் ஜார்ஜ் டவுன் பல்கலைக் கழகம் 850 ரஷ்ய சொற்களையும் இலக்கணத்தின் ஆறு விதிகளையும் கொண்டு ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை நடைமுறைபடுத்திக் காட்டியது. அடிப்படைச் சிக்கல் பொறியியல் தன்மை அடிப்படையிலானது என்று நம்பப்பட்டது. அடுத்த பத்தாண்டுகளில் USIA இல் இத்தகைய ஆய்வுகள் பெருமளவில் ஊக்குவிக்கப்பட்டன. கடினமான மொழியியல் சிக்கல்கள் குறித்த விழிப்புணர்வு அதிகரித்தது. திருப்தியில்லாத முன்னேற்ற காரணமாக ALTPA இயந்திர மொழிபெயர்ப்பிற்கானப் பண உதவியைக் குறைக்கப்பரிந்துரைத்தது.

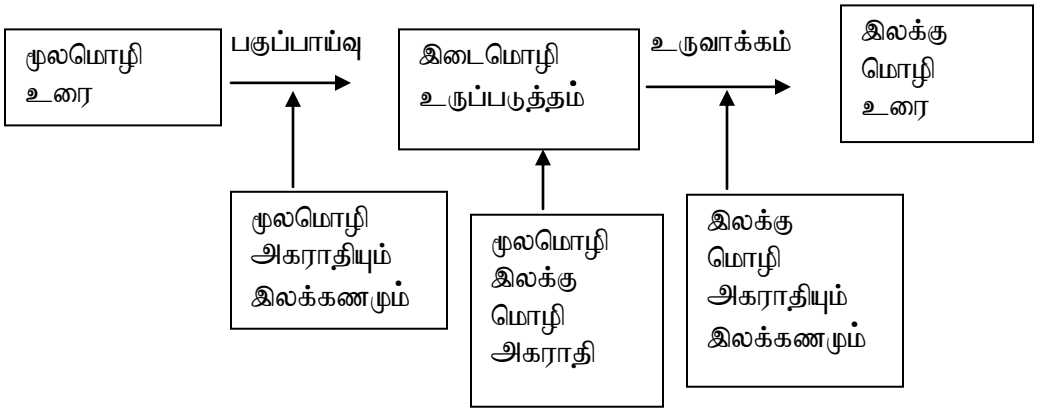
இக்காலகட்டத்தில் தொடரியல் ஆய்வு தொடக்க நிலையில் இருந்தது. இலக்கண விதிகள் மற்றும் தொடரியல் அமைப்பு ஆகிய கருத்துச சாயல்கள்

இல்லாதிருந்தது ஆங்கிலம் மற்றும் ருஷ்ய இலக்கணங்கள் வழியமைப்பின் (program) அமைப்பின் ஒரு பகுதியாக இருந்தது.



மூன்றாவது காலகட்டம் (1966-1975)

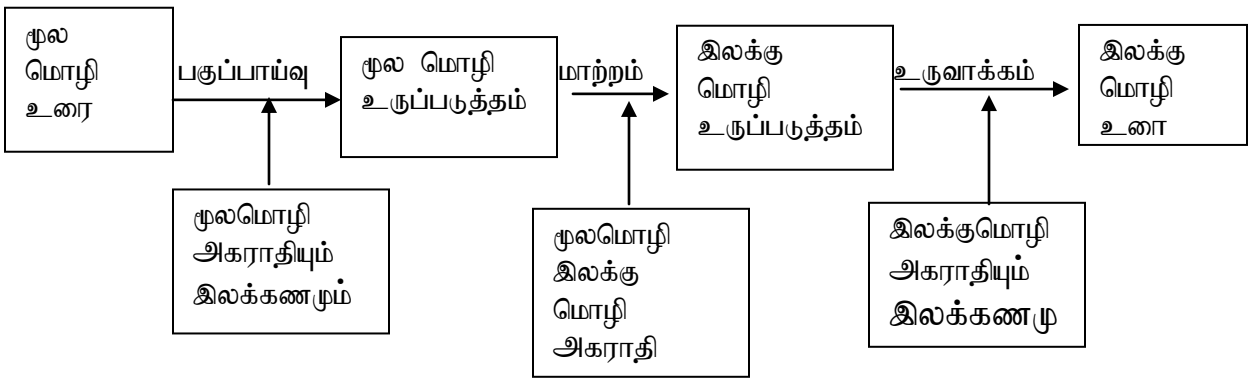
நேரடியான மொழிபெயர்ப்பு அடிப்படையில் இயங்கும் இரு ஒழுங்கமைப்புகள் System மற்றும் Logos. இந்தக் காலகட்டத்தில் மொழியியல் கோட்பாட்டின் முன்னேற்றத்தில் மிகுந்த மாற்றம் ஏற்பட்டது. மொழியியல் தரவு கணிப்பொறியின் செயல்பாட்டிலிருந்துப் பிரிக்கப்பட்டது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் இடைமொழி அணுகுமுறையைப் (interlingua approach) பயன்படுத்துகிறது; இதன் செயல்பாட்டின் இரண்டு நிலைகள் இருந்தன. மூலமொழியிலிருந்து இடைமொழிக்கு மற்றம் இடைமொழியிலிருந்தும் இலக்கு மொழிக்கு.



Metal மற்றும் Ceta என்ற இரண்டு ஒழுங்கு முறைகள் இடைமொழி அணுகுமுறையின் (interlingua approach) அடிப்படையில் உருவாக்கப்பட்டவை; இடைமொழி தொடரியல் அமைப்பிற்கு எல்லைப்படுத்தப்பட்டது. பொருண்மை முலகூறுகளாகச் சொற்களைப் பிரிப்பதற்கு முயற்சிகள் மேற்கொள்ளப்படவில்லை.

இந்த காலகட்டத்தில் உருவாக்கிய மற்றொரு மாதிரி மாற்ற அணுகுமுறையாகும். (Transfer Approach) இந்த அணுகுமுறையில் மூலமொழிக்கும் இலக்கு மொழிக்கும் அவைகளுடைய தனித்தன்மையான அக அமைப்புகளின் உருப்படுத்தம் (deep structure representation) இருக்கும். மொழிபெயர்ப்பு மூன்றுமட்ட செயல்முறையாகும்.

1. உரைகளை மூலமொழியின் உருப்படுத்தமாகச் செய்தல்,
2. அதை இலக்குமொழி உருப்படுத்தமாக மாற்றுதல்,
3. இலக்கு மொழியின் உரைகளை உருவாக்குதல்,



மாற்ற ஒழுங்குமுறையின் மிக அறியப்பட்ட எடுத்துக்காட்டு மாண்ட்ரியல் பல்கலைக் கழகத்தில் செய்யப்பட்ட ஆங்கில பிரெஞ்சு மொழி ஒழுங்கு முறையான TAUM திட்டம் ஆகும்.

நான்காவது காலகட்டம் (1975-1985)

இந்தக் காலகட்டத்தில் உருவாக்கப்பட்ட ஒழுங்குமுறை குவிதல் அணுகுமுறைகளை (convergence appporach) பயன்படுத்தியது. இவைகள் கட்டுப்படுத்தப்படாத உரை உள்ளீடு கொண்ட முழு தானியங்கு ஒழுங்குமுறைகளின் திட்ட அமைப்பிற்கு எல்லைக்குட்பட்ட மொழி ஒழுங்குமுறைகளும் ஊடாடும் ஒழுங்குமுறைகளும் உருவாக்கப்பட்டன. தொடரியல் மற்றும் பொருண்மையியலுக்கு எல்லைப்படுத்தப்பட்ட ஒழுங்கு முறையின் எடுத்துக்காட்டு TAUM என்பதன் அனுபவத்தின் அடிப்படையில் மாண்டரியல் பல்கலைக் கழகத்தால் உருவாக்கப்பட்ட Meteo என்பதாகும் ; இது ஆங்கிலத்தில் உள்ள வானிலை அறிக்கைக்களைப் பிரெஞ்சு மொழிக்கு மொழிப்பெயர்க்கும். ஊடாட்டம் செய்யும் இயந்திரமொழி பெயர்ப்பின் கவர்ச்சி மனித மொழிப்பெயர்ப்பாளர்கள் மற்றும் கணினி கணிப்பொறிகள் இவற்றின் ஒத்துழைப்பைச் சிறப்பாகப் பயன்படுத்துவதில் உள்ளது.

இன்றைய காலகட்டம் (1985-இல் இருந்து)

கட்டுப்படுத்தும் மற்றும் இணைக்கும் கோட்பாடு (Government and Building (GB)) Theory), சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), கிளை இணைக்கும் இலக்கணம் (Tree Adjoining Grammar (TAG)), பெரிதாக்கப்பட்ட மாற்ற வலைப்பின்னல் (Augmented Transition Net work (ATN)) போன்ற புதிய இலக்கண

வடிவங்களின் வருகை இயற்கை மொழி ஆய்வுப் பரப்பில் புதிய பார்வையை உருவாக்கியது. இந்த மொழியின் வடிவமைப்புகள் பகுத்துக் குறிக்கும் வழிமுறைகளில் தொடரியல் மற்றும் பொருண்மையியல் பண்புக் கூறுகளின் பொருத்தத்திற்கு ஆலோசனைக்குரியது. மேலும் கணிப்பொறி அறிவியலில் ஏற்பட்ட தொழில்நுட்ப முன்னேற்றம் இயந்திர மொழிபெயர்ப்பு ஆய்வின் வளர்ச்சிக்குச் சாதகமாக அமைந்தது. சொல்-பொருண்மையில் பகுப்பாய்வை உள்ளடக்கிய இயற்கை மொழியின் புரிதல் (Natural Language Understanding(NLU) தேவையான உலக அறிவை அடையானம் காணுவதை உள்ளடக்கியது. அறிவு உருப்படுத்தம் மற்றும் மொழியியல் அறிவில் உள்ளடக்குவது போன்றவை இயந்திர மொழிபெயர்ப்பு ஆய்வுகளின் பரப்பில் மேலும் கூடுதல் வளர்ச்சிக்கு உதவி புரிந்தது. தற்கால ஒழுங்குமுறைகள் இந்த மூன்று அணுகுமுறைகளின் மேம்பாடுகளை பயன்படுத்துகின்றன. முழு ஒழுங்குமுறையும் பல எண்ணிக்கையிலான துணை ஒழுங்குமுறைகளாக பிரிக்கப்பட்டுச் சரியாக இடைமுகப்படுத்தப்பட்டுள்ளன. இயந்திர மொழிபெயர்ப்பின் இவ்வேறுபட்ட கட்டங்களை மூன்று நிலைகளாகக் குழுமலாம்:

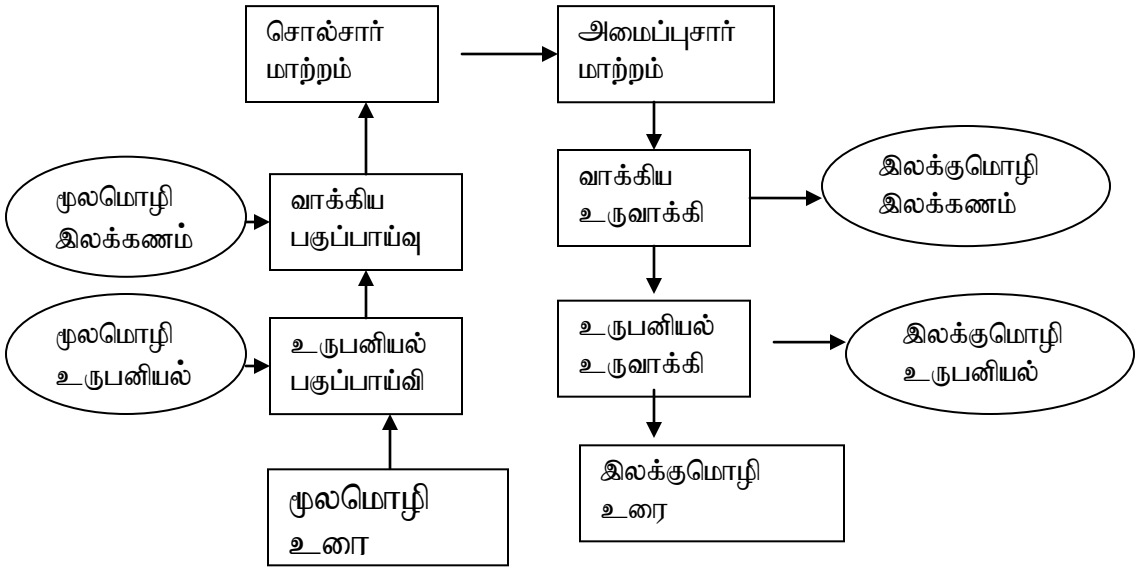
1. **ஆய்வு நிலை:** ஒழுங்குமுறையின் உள்ளீடாக மூலமொழி உரை முதலில் சொல் நிலையில் ஆயப்படுகிறது. உருபனியல் ஆய்வு உருபனியல் விதிகள், சந்தி விதிகள், அடிச்சொல் இயந்திரம் படிக்கவியலும் அகராதி மற்றும் முன்னொட்டு இயந்திரம் படிக்கவியலும் அகராதி போன்ற பலதரப்பட்ட அகராதிகளைப் பயன்படுத்தி சொற்களை அவற்றின் உருபன்களுக்காகப் பிரிக்கிறது. வாக்கியப் பகுப்பாய்வியல் ஒரு இலக்கண மாதிரியைப் பயன்படுத்தி ஒவ்வொரு வாக்கியத்தையும் அவற்றின் உறுப்புக்களான தொடர்கள் மற்றும் எச்சத் தொடர்களுக்காகப் பகுப்பாய்வு செய்கிறது; பயனிலைகளுக்கும் அவற்றின்

பங்கெடுப்பாளர்களுக்கும் இடையில் உள்ள தொடரியல் மற்றும் பொருண்மையியல் உறவுகள் குறியாக்கம் செய்யப்படுகின்றன.

2.**மாற்ற நிலை** : இந்த நிலையில் இரண்டு கட்டங்கள் இருக்கின்றன. 1. சொல் சார் மாற்றம் 2. அமைப்பு சார் மாற்றம். சொல் சார் மாற்ற நிலையில் உரை மற்றும் அகராதியின் தொடரியல் மற்றும் பொருண்மையியல் குறியாக்கத்தின் உதவியுடன் ஒவ்வொரு சொல்லின் சொல் சார் பொருண்மையின் மெய்ப்படுத்தம் செய்யப்படும்; இலக்கு மொழியிலிருந்து பொருத்தமான சொல் தேர்ந்தெடுக்கப்படும்.

இந்த அமைப்பு மாற்றக் கட்டத்தில் இலக்குமொழி, மூலமொழி விதிகளைப் பயன்படுத்தி மூலமொழி அமைப்பு இலக்குமொழி அமைப்பிற்கு மாற்றப்படும்.

3. உருவாக்க கட்டம்: இந்த கட்டத்தில் வாக்கிய உருவாக்கும் பகுதி (Sentence Generator Module) இலக்குமொழி இலக்கணத்தின் உதவியால் பெறப்படும் அமைப்புகளுக்கு இலக்கண அடிப்படையில் சரியான வாக்கியங்களை உருவாக்கும். பின்னர் உருபனியல் விதிகள் அல்லது புணர்ச்சி விதிகள், இயந்திரம் படிக்கவியலும் அடிச்சொல் அகராதி, இயந்திரம் படிக்கவியலும் முன்னொட்டு அகராதி இவற்றைப் பயன்படுத்தி உருழூனியல் சார் குறியாக்கம் இலக்குமொழியின் பொருத்தமான ஒட்டுக்களால் இடம் பெயர்க்கப்படும்; வெளியீடு இலக்குமொழியில் ஒரு உரையாக உற்பத்தி செய்யப்படும். இந்த மொழிபெயர்ப்பு ஒழுங்குமுறைகளும் வேறுபட்ட நிலைகளும் (மட்டங்களும்) அவற்றின் துணை ஒழுங்குமுறைகளும் கீழே தரப்பட்டுள்ளன.



கணிப்பொறி வழி மொழபெயர்ப்பினை தானியங்கி மொழிபெயர்ப்பு என்றும், இயந்திர மொழி பெயர்ப்பு என்றும், கணிப்பொறி உதவியுடன் நடைபெறும் மொழி பெயர்ப்பு என்றும் கொள்ளலாம். கணிப்பொறியின் வேகம், நினைவகத் திறன், பதிவு செய்துள்ள செய்திகளை தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளபெறக் கூடிய வசதி போன்றன கணிப்பொறி வழி மொழி பெயர்ப்பு பணியை மேற்கொள்ள உதவின.

கணிப்பொறி வழி மொழி பெயர்ப்பு பணி 1950 –ம் ஆண்டு வாரன் வீவர் (Warren Weiver) என்ற அறிஞரால் தொடங்கப்பெற்றது. இதைத் தொடர்ந்து ஓ பார் ஹில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்பு பணியை ஆரம்பித்தார். 1954-ம் ஆண்டு ஜார்ஜ் டவுன் பல்கலைக் கழகம் கணிப்பொறி வழி மொழி பெயர்ப்பு பணியை துவங்கியது. இதைத் தொடர்ந்து 1956-ம் ஆண்டு மாஸ்கோவில் பிரெஞ்சு மொழிகளிலிருந்து ரஷ்ய மொழிக்கு கணிப்பொறி வழி மொழிபெயர்ப்பு பணி

மேற்கொள்ளப்பட்டது. இப்பணியை மேல் ஜங்க் (Melchunk) என்ற அறிஞர் மேற்கொண்டார்.

1973-ம் ஆண்டு அமெரிக்காவில் உள்ள LOGOS Incorporation Limited நிறுவனம் மொழி பெயர்ப்பிற்கான மென்பொருளை வடிவமைத்தது. இம்பென்பொருளைக் கொண்டு ஆங்கிலத்தில் அமைந்துள்ள செய்திகளை வியட்நாம் மொழியில் மொழி பெயர்ப்பு செய்யப்பட்டது. ஏறக்குறைய இதே காலக் கட்டத்தில் Tertils என்ற நிறுவனம் TITUS என்ற மென்பொருளை வடிவமைத்தது. இம்மென்பொருளைக் கொண்டு ஆங்கிலம், ஸ்பானிஷ், பிரெஞ்ச் மற்றும் ஜெர்மனி ஆகிய மொழியில் உள்ள செய்திகள் ஒனிறிலிருந்து மற்றவைகளுக்கு மொழி மாற்றம் செய்யப்பட்டது. இதைத் தொடர்ந்து CUTD CHINESE UNIVERSITY TRANSLATOR என்ற மென்பொருள் உருவாக்கப்பட்டது. இதைக் கொண்டு சீன மொழியில் உள்ள செய்தித் தாள்கள் ஆங்கில மொழிக்கு மொழி மாற்றம் செய்யப்பட்டன.

1974-ம் ஆண்டு SYSTRAN என்ற மென்பொருள் உருவாக்கப்பட்டது. SYSTRAN என்பதன் விரிவு System Translation என்பதாகும். இம்பொழி பெயர்ப்பு மென்பொருள் உருஷ்ய மொழியிலிருந்து பிரெஞ்ச் மொழிக்கு மொழி மாற்றம் செய்யப்பட்டது. இதற்கானத் திட்டத்தை P. TOMA என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன் முதலில் வெளியிடப்பட்ட கணிப்பொறி வழி மொழி பெயர்ப்பிற்கான மென்பொருள் இதுவேயாகும்.

1975-ம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென்பொருள் உருஷ்ய மொழியில் அமைந்துள்ள செய்திகளை பிரெஞ்ச் மொழிக்கு மாற்றம் செய்யப்படுவது. இதே காலக் கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு மொழிபெயர்க்க வடிவமைக்கப்பட்டது.

இதுபோன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனை கட்ட முயற்சியாகவும், வணிக நோக்கத்திலும் கணிப்பொறி வழி மொழி பெயர்ப்பிற்கான மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைப்பெற்று வருகின்றன. 1985-ம் ஆண்டு தமிழ்பபல்கலைக் கழகத்தில் TUMTS (Tamil University Translation System) என்ற திட்டப்பணி தொடங்கப் பெற்றது. உருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிபெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக் கழகங்களில் கணிப்பொறி அறிவியல் துறை, மொழியில் துறை மற்றும் மொழிபெயர்ப்பு துறை ஆகியவைகளின் கூட்டு முயற்சியில் விளைந்ததாகும்.

இதற்குப் பின்னால் கான்பூரில் உள்ள இந்திய தொழில் நுட்ட நிறுவனம் (Indian Institute of Technology) ஹிந்தியிலிருந்து ஆங்கிலம், ஆங்கிலத்திலிருந்து ஹிந்தியை மொழி பெயர்க்கவும் இந்தியிலிருந்து(தமிழ், தெலுங்கு, கன்னடா போன்ற) இந்திய மொழிகளுக்கு திட்டமிட்டு செயலாற்றியது. இப்போது ஹைதராத்தில் உள்ள உலகத் தகல் தொழில் நுட்பநிறுவனம் (Indian Institute of Information Technology (IIIT), பத் முழுமுயற்சியுடன் இயந்திர மொழி பெயர்ப்பில் ஈடுபட்டுள்ளது. தற்பொழுது சென்னையிலுள்ள AU- Anna University , KRC- Kalyana Sundharam Reasch Center என்ற ஆய்வு நிறுவனம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழி பெயர்க்க திட்டம் ஒன்றினை மேற்கொண்டு செயல்படுத்தி வருகிறது.

சமீபத்தில் கருத்துப்பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் தகல் தொழில் நுட்ப நிறுவனத்தின் நிதி நல்கையில் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கும் மொழிபெயர்ப்பு செய்ய வேண்டி இரண்டு குழுக்கள் (consortium) அமைக்கப்பட்டு செயல்பட உள்ளன. இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்க்கும் குழுவின் தலைவராக பேராசிரியர் இராஜிவ் சங்கல் அவர்கள் செயலாற்றுகின்றார். இக்குழுவின்

அங்கத்தினாகளின் ஒருவராக தமிழ்ப்பல்கலைக்கழக மொழியில் துறை பேராசிரியர் ச.இராசேந்திரன் அவர்கள் இருக்கின்றார். இவரிடம் மலையளம்-தமிழ் மொழிபெயர்ப்பு ஒப்படைக்கப்பட்டுள்ளது.

2.18.3. கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிகளுக்கு, மொழிமாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்தகைய கணிப்பொறி வழி மொழிபெயர்ப்பு பணிகளில் பின்வரும் இரு எல்லைகள் நிர்ணயிக்கப் படுகின்றன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் வட்டார அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக் கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்துவது என்பதற்கோ, அதே பொருளில் பயன்படுத்துவது என்பதற்கோ உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழியை இலக்கிய மொழி, பேச்சு மொழி என இரு வகையாகப் பிரிக்கப்படுகிறது. பேச்சு மொழியை கணிப்பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிபெயர்ப்பு பணிகளில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக் கொள்ளப்படுகின்றன.

2. செய்யுள் கவிதை போன்றவைகளில் பொருளுக்கு தரும் முக்கியத்துவத்தை விட கருத்தினைப் புலப்படுத்தும் சக்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகினறன. இவைகள் மனித உளவியலை பிரதிபலிப்பான இவ் உளவியல் கூறுகளை கணிப்பொறி வழி வெளிப்படுத்துவதென்பது இயலாத ஒன்றாகும். இதுபோன்று இலக்கியங்களிலும் பொருளை விட உளவியல் கூறுகளுக்கு

முக்கத்துவங்கள் தரப்படுகின்றன. எனவே செய்யுள், கவிதை, பாடல்கள் மற்றும் இலக்கிய செய்திகள் ஆகியவைகளை கணிப்பொறிற வழி மொழிப்பெயர்ப்பிற்கு தொடக்க நிலையில் எடுத்துக் கொள்ளப்பட வில்லை. அறிவியல் தொழில் நுட்பம் சார்ந்த படைப்புகளில் பொருளுக்கே முக்கியத்துவம் தரப்படுவதில்லை. இங்கு கோபம், வீரம், தாபம், எழுச்சி போன்ற உணர்ச்சிகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழப்பெயர்ப்பு பணிகளுக்கு பொதுவாக அறிவியல் மற்றும் தொழில் நுட்பத்தை சார்ந்த படைப்புகளே எடுத்துக் கொள்ளப் படுகின்றன. மேற்கூறிய இரண்டு கணிப்பொறி வழி மொழி பெயர்ப்பிற்கு எல்லைகளாக கொள்ளப்படுகின்றன.

2.18.4. உருஷ்யன்-தமிழ் மொழி பெயர்ப்பு ஒரு கண்ணோட்டம்

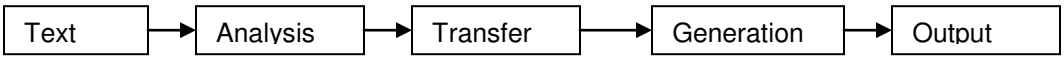
இரு மொழிகளுக்கு இடையே உள்ள சொல்வளம் (vocabulary), இலக்கணம் (grammar), கருத்தாடல் (context Structure) இவற்றின் ஒற்றுமை வேற்றுமைகளை ஆராய்வது வகைப்பாடடியல் ஆகும். இரு மொழிகளுக்கு இடையே உள்ள வேற்றுமைகளை மட்டும் அறிந்து ஆராய்வது வேற்றுமை பாராட்டியாகும். உருஷ்யன்-தமிழ் மொழிபெயர்ப்பு பணியில் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமைகள் ஆராயப்பட்டு தொடங்கப் பெற்றது. இவ்விரு மொழிகளுக்கு இடையே உள்ள சொல்வளம், சொற்களுக்கு இடையே உள்ள உறவுகள், தொடரியல் மற்றும் வாக்கியங்களுக்கு இடையே உள்ள உறவுகள் கண்டறியப்பட்டு வரைமுறைபடுத்தப்படுகின்றன. இம்மொழி பெயர்ப்பிற்கு பின்வரும் இரு குறிகோள்கள் அடிப்படையாகக் கொள்ளப்பட்டன:

1. எழுத்து அல்லது இலக்கிய உரைகளை மட்டும் மொழி பெயர்ப்பிற்கு எடுத்துக் கொள்வது. பேச்சு மொழி வழக்கினை ஒதுக்குவது

2. இலக்கிய படைப்புகளை தவிர்த்து அறிவியல் மற்றும் தொழில் நுட்ப உரைகளை மட்டும் எடுத்துக் கொள்வது

2.18.4.1. முன்னேற்றப் பணிகள்

இம்மொழி பெயர்ப்புக்கு உருஷ்ய மொழி மூலமொழியாகவும் (Source Language) தமிழ் இலக்கு மொழியாகவும் (Target Language) எடுத்துக் கொள்ளப்பட்டன. மூலமொழி அதாவது ருஷ்யன் மொழியில் அமைந்துள்ள செய்திகளை ஒலிபெயர்ப்பு (translitration) செய்து கணிப்பொறிக்கு உள்ளீடு செய்வது. இம்மொழி பெயர்ப்புக்கு உருஷ்ய மொழியை மூலமொழியாகவும் (Source Language) தமிழை இலக்கு மொழியாகவும் (Target Language) எடுத்துக் கொள்ளப்பட்டன. உருசிய மொழிக்கும், தமிழுக்கும் ஒரு அகராதி உருவாக்கப்பட்டது. உருஷிய மொழியில் அமைந்துள்ள சொற்கள் சொறறொடர்கள் மற்றும் வாக்கிய அமைப்புகளில் சில முன்திருத்தம் (Preprocessing) செய்தல். மூல மொழியில் உள்ள சொல் மற்றும் வாக்கிய அமைப்பிற்கு நிகரான தமிழ்ச்சொல் வாக்கியம் ஆகியவற்றை நிர்ணயம் செய்ய விதிமுறைகளை உருவாக்குவது இத்தகைய விதிமுறைகளும், வழிமுறைகளும் மாற்றம் (Transfer) என்ற பெயரில் பதிவு செய்தல். கணிப்பொறி வழி மொழி பெயர்ப்பு செய்த பின்னர் தேவைப்படின் தமிழ் மொழிக்கேற்ப வாக்கிய அமைப்புகளை திருத்தி அமைப்பது (Post Processing).



உள்ளீடு பகுப்பாய்வு மொழிமாற்றம் உருவாக்கம் வெளியீடு

2.18.4.2 உருசியன்-தமிழ் மொழி பெயர்ப்பின் படிநிலைகள்

உருஷ்யன்-தமிழ் மொழி பெயர்ப்பிற்கான படிநிலைகளின் விளக்கம் கீழே தரப்பட்டுள்ளன.

மூலமொழி உருசியனை L1 என்றும், தமிழ் மொழியை L2 என்றும் குறிப்பிட்படுகின்றன. முதற்கட்டமாக மூலமொழியில் உள்ள செய்தி ஒலிபெயர்ப்பு செய்யப்பட்டு கணிப்பொறிக்கு உள்ளீடு செய்யப்படுகின்றன. 2-ஆம் கட்ட நிலையில் மூலமொழியில் உள்ள செய்தி பகுப்பாய்வு செய்யப்படுகிறது. இந்நிலையில் சொற்களின் இலக்கணக் கூறுகள் கண்டறியப் படுகின்றன. இதற்கு உருஷ்யன்தமிழ் அகராதி உதவுகிறது. இப்பகுப்பாய்வு மூலம் ருஷ்ய மொழியில் உள்ள செய்திகள் NP, UP, N.N, ADI எனப் பகுப்படுத்தப்படுகின்றன.

மூன்றாம் கட்ட நிலையில் பகுப்பாய்வு செய்யப்பட்டுள்ள மூலமொழியின் வார்த்தையின் சொற்களின் வடிவம், வாக்கிய அமைப்பு போன்றவைகளுக்கு இலக்கு மொழியில் நிகராக உள்ள சொற்களின் வடிவம் வாக்கிய அமைப்புன் மூலம் கண்டறியப்பட்டுள்ளது. மொழி மாற்றம் செய்யப் படுபகின்றன. இந்நிலையில் இவ்விரு மொழிகளுக்கும் இடையே உள்ள முரண் அறிவு (Contrast Knowledge) பயன்படுத்தப்படுகின்றன. கருத்தாடலை பொறுத்த வரையில் இவ்விரு மொழிகளுக்கு இடையே ஒன்றுக்கு-ஒன்று என்ற (one–to–one) உறவு உள்ளது. வேற்றுமைகள் அதிகரிக்க அதிகரிக்க சிக்கல்கள் ஏற்படுகின்றன. பெயர் வரிச் சொல்லைப் பொறுத்த வரையில் இவ்விரு மொழிகளுக்கு இடையே பலவற்றிற்கு-ஒன்று என்ற (many–to–one) உறவு உள்ளது. இவ்வுறவினை கீழ்க்கண்ட படம் விளக்குகிறது.

உருசியன்	தமிழ்
ADJ	
SN	
PL	ADJ
CASES	
GENDERS	

மேலும் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமை உறுப்புகளில் ஒன்றிற்கு பல என்ற (one-to-many) உறவு உள்ளது.

உருசியன்	தமிழ்
	KIRAAL – கின்றாள்
ON	KIRAAN – கின்றான்
	KIRAAR – கின்றார்

இது போன்று இவ்விரு மொழிகளுக்கு இடையே உள்ள வேறுபாடுகள் விதிமுறைகளாகவும் வழிமுறையாகவும் மாற்றம் (transfer) என்ற அமைப்பின் கீழ் தரப்பட்டுள்ளது. பகுப்பாய்வு செய்யப்பட்ட மூலமொழி செய்தி அமைப்பின் உள்ளீடாக செல்கிறது. இந்நிலையில் மூலமொழி அமைப்பிற்கு நிகரான இலக்கு மொழி அமைப்பிற்கு செய்தி மொழி மாற்றம் செய்யப்படுகிறது.

நான்காவது கட்ட நிலையில் மொழி மாற்றம் செய்யப்பட்டுள்ள இலக்குமொழி கட்டமைப்புகள் இலக்கு மொழியின் தன்மைக்கேற்ப அமைவதில்லை. இவ்வாறு கிடைக்கப்பெற்ற இடைநிலை மொழி மாற்ற அமைப்புகளை இலக்குமொழியின் தன்மைக்கேற்ப உருவாக்கி தருவது தான் GENARATION என்ற பகுதியின் பணி ஆகும். இநிலையில் மொழிமாற்றம் செய்யப்பட்ட செய்திகள் உள்ளன. இச்செய்தியை மூலமொழியின் வரிவடித்திலோ அல்லது ஒலிபெயர்ப்பு வடிவத்திலோ தரும் பணியை

output என்ற அமைப்பு தருகிறது. மொழிபெயர்ப்பு செய்யப்பட்ட செய்தியைக் காட்சித் திரையில் காணலாம். உருஷியன்-தமிழ்மொழி பெயர்ப்பைப் பொறுத்த வரையில் வெளியீட்டை தமிழ் எழுத்துகளில் காண்பதற்கும் அச்சிட்டுக் கொள்வதற்கும் வசதிகள் உள்ளன. அதுபோல் ஒலிபெயர்ப்பு முறையில் விடையை பெறலாம்.

2.18.5. கணிப்பொறியின் மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழி பெயர்ப்பு பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற போது கணிப்பொறி சிறப்புத் தன்மைகளான அதிவேகம், மிகுந்த நினைவகக் கொள்திறன், தகவல்தளம் உருவாக்கும் வசதிகள் போன்றவைத் திறமையாகப் பயன்படுத்தப்படுகின்றன. ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி பெயர்ப்பு செய்யப்படும்பொழுது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வடிவமைக்கப்படுகின்றன. எனவே இருமொழி மொழிபெயர்ப்பு பணிக்கு கூட ஒருவருக்கு இருமொழிகளிலும் திறமை இருக்க வேண்டும் என்ற அவசியம் இல்லை. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணிகளில் பகுப்பாய்வு சிக்கல்கள் உள்ளன. ஒரு மொழியின் வளம், தன்மை ஆகியவற்றை கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையானப் பணி அல்ல. எனவே கணிப்பொறி வழி மொழிபெயர்ப்பு திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப் பெற்றிருப்பினும் இன்னும் முழுமைப் பெறவில்லை. மேலும் கணிப்பொறி அறிவியலின் தொழில் நுட்ப வளர்ச்சி கணிப்பொறியில் மொழி பெயர்ப்பு பணிக்கு மிகவும் பயனுள்ளதாக அமைகிறது. இத்தகைய பணிகளின் குறை நிறைகளை கீழே காண்போம்.

மொழிகளுக்கு இடையே உள்ள சொல் வளம் (வடிவம்), இலக்கணம், கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிகளுக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிப்பெயர்ப்பு பணியின் சிக்கல்கள் குறையும்.

2.18.5.1. உருபனியல் சிக்கல்

உருபனியல் அமைப்பை பொறுத்த வரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

1. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து
2. உருபன்களின் கட்டமைப்பொறுத்து

ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் மட்டும் இருக்குமேயானால் அவைகளை எவ்வளவு துல்லயமாகப் பிரிக்க இயலும் என்பதை பொறுத்து ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தால் அந்த மொழியை தனிநிலை உருபன் கட்டு (isolating) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டு (polysynthetic) மொழி என்கிறோம். சிபேரியன் மொழி பன்நிலை உருபன் மொழிக்கு சிறந்த உதாரணம் ஆகும். வியட்நாமில் மற்றும் கேண்டுனஸ் ஆகிய மொழி தனிநிலை உருபன் கட்டு மொழிக்கு சிறந்த உதாரணம் ஆகும். எனவே இத்தகைய மொழிகளுக்குள் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்றபொழுது மொழிகளின் உருபனியல் பற்றிய விரிவான தகவல் அவசியமாகும்.

2.18.5.2. தொடரமைப்புச் சிக்கல்

மொழிகளின் தொடரியல் அமைப்பை (Syntactic Structure) பொறுத்து மொழிகள் என மூன்று வகைகளாகப் பிரிக்கப்படுகின்றன:

1. SVO அமைப்பு

ஜெர்மன், பிரெஞ்ச், ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் இந்த அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரிஸ், அரேபி, ஹிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருபுகள் (preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிப் பெயர்ப்புத் திட்டங்களுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரியல் *வீடு* என்பது தலைச்சொல் (head word) ஆகும். *இராமன்* என்பது சார்பு சொல் ஆகும். தலைச் சொல்லுக்கும், சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை *இன்* என்ற ஒட்டு (affix) விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச்சொல்லுக்கும் இடையே உள்ள உறவை வெளிப்படுத்தும் முறையைப் பொறுத்து மொழியைத் தலைக் குறித்தல் மொழி (head marking languge) சார்புக் குறித்தல் மொழி (dependent marking language) என இரு வகையாகப் பிரிக்கப்படுகிறது. தலைக்குறித்தல் மொழியின் ஒட்டு சார்புச்

சொல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி சார்புக் குறித்தல் மொழியை சார்ந்ததாகும். ஹங்கேரிய மொழி தலைக்குறித்தல் மொழியை சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு இதை விளக்கும்:

ENGLISH	: Man's house
Hungarian	: The man house his
H	: Head word (தலைச்சொல்)
The Man	: Dependent Word (சார்புச்சொல்)
A	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணியை மேற்கொள்கின்றபோது தலைச்சொல்லுக்கும் சார்புச்சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப்பெறுகின்றன என்பது பற்றியெல்லாம் ஆராய வேண்டும்.

2.18.5.3. சொல்வகைப்பாடு பொருத்தச் சிக்கல்

மூலமொழியிலுள்ள ஒரு சொல்லின் வகைபாடும் (part-of-speech) அச்சொல்லுக்கு நிகராக இலக்கு மொழியலுள்ள சொல்லின் வகைபாடும் ஒன்றாக இருக்க வேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக, ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழிப்பெயர்ப்பு செய்யப்படும்போது மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்க வேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்பு பணியில் உள்ள சிக்கல்கள் ஆகும்.

2.18.5.4. கருத்தாடல் சிக்கல்

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கான தளம் விளக்கமானதாகவோ சுருக்கமானதாகவோ இருக்கலாம். எடுத்துக்காட்டாக, சகோதரன் என்ற உறவைக்

குறிக்க ஆங்கிலத்தில் brother என்ற சொல் உள்ளது. மூத்த சசோதரனை elder brother என்றும் சொல்கிறோம். இவ்வாறு வயதில் மூத்தவர், இளையவர் என்பதைக் குறிக்க elder, younger என்ற சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் அண்ணன் என்றும், தம்பி என்றும் தனித்தனிச் சொற்கள் உள்ளன. இங்கு சகோதரன் என்ற உறவு மேலும் வகைப்படுத்தப்பட்டு முதியவர், இளையவர் என்ற வகையில் அண்ணன், தம்பி என்ற இரு சொற்கள் உள்ளன. ஆனால் அது ஆங்கிலத்தில் இவ்வாறான வகை என்பது இல்லை. உறவை வேறுப்படுத்த elder, younger என்ற சொற்கள் பயன்படுத்த நேரிடுகிறது.

2.18.5.5. கலாச்சார வேறுபாட்டுச் சிக்கல்

மூலமொழியிலுள்ள செய்திகளின் கலாச்சாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்குமொழிக்கும் பொருத்தமானதாக அமைதல் வேண்டும் அல்லது இலக்கு கலாச்சார பண்புகளுக்கு ஏற்ப மாற்றித் தக்க அளவு இருத்தல் வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சாரத் தொடர்புகள் மொழிபெயர்ப்புச் செய்ய இயலுமா?, இயலாத? என்பதைத் தெளிவு படுத்தும் . மேலும் சபீர்-வீ.ப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ அல்லது குறைப்பதாகவோ உள்ளது. எனவே மொழிப்பெயர்ப்பில் இத்தகைய உளவியல் ரீதியான சிக்கல்கள் உள்ளன.

2.18.5.6. தொடரியல் பகுப்பாய்வுச் சிக்கல்

தொடரியல் பகுப்பாய்வு (parser) பலப் பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிப்பெயர்ப்பிற்காக பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்கு தொடரின் அல்லது வாக்கியத்தின்

பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வுச் செய்யப்படுகிறது. எனவே இயந்திர மொழிப்பெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது. பொருளை சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அதிக அடிப்படை (Knowledge Base) தேவைப்படுகிறது.

2.18.5.7. அகராதிச் சிக்கல்

மொழிப்பெயர்ப்பு பணிகளுக்காக உருவாக்கப் படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிற வகை அகராதிகளை விட துறைச் சார்ந்த அகராதிகள் மொழிப்பெயர்ப்பு பணியை எளிமைப் படுத்துகிறது. மூல மொழியிலுள்ள செய்தியை பொறுத்து அதன் தொடர்புடைய அகராதியைப் பயன்படுத்திக் கொள்ளலாம். மொழிப்பெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத்தொடர்கள், மேலும் அடிக்கடி பயன்படுத்தக் கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்பட வேண்டும்.

2.18.5.8. பிற சிக்கல்கள்

இயந்திர வழி மொழிபெயர்ப்பு செய்கின்றபொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்கு கூடுதல் தகவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கல்களைப் புரிந்துக் கொண்டு அதற்கேற்ற தீர்வு உரிய முறையில் வழங்கப்பட வேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்க வேண்டியது அவசியமாகும். பெரும்பாலும் படைப்பாற்றல் திறன்

செய்திகளில் வெளிப்படுத்தும்பொழுது கணிப்பொறிக்கு கூடுதல் செய்திகள் தேவைப்படுகின்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிப்பெயர்ப்புகளில் உள்ளீடு செய்திகள் முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும், இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. இதை இருவகைகளில் செய்யலாம்:

அ) மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிப்பெயர்ப்பது. பின்னர் மொழிப்பெயர்ப்பை படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகிறதா? என்பதை ஆராய்வது, அப்படி இல்லை மூலமொழிச் செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிப்பெயர்ப்பு செய்வது சரியான கருத்து வெளிவரும் வரை இத்தகைய ஆய்வை மீண்டும் செய்வது.

ஆ) மொழிப்பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து தேவையானத் திருத்தங்களை செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தையும் மிச்சப்படுத்துகிறது. மொழிப்பெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிப்பெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப் பணி திறமையானதாக அமைகிறது.

கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிஎன்பது ஒரே முச்சில் ஒட்டு மொத்தமாகச் செய்யக் கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்பட வேண்டும். அதன் செயல் திறனை அறிய வேண்டும். பணியின் தன்மையை மதிப்பிட வேண்டும். இதற்கு ஏற்றாற் போல் தேவைப்படின் மென்பொருள் கூடுதல் வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை

மனிதனுக்கு நிறைவழிக்கும் செயல்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்பு கடுமையானப் பணியாகும். இதற்கு மனித மொழிப்பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (models) வழி முறைகளும் அவசியமாகும். கணிப்பொறி மொழிப்பெயர்ப்பு எண்ணற்ற முயற்சிகளினால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமானப் பணியாகும். இங்கு அறிவியல் மட்டுமல்லாது மொழிப்பெயர்ப்பிற்கு தேவையான கலைத் திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் உருவாக்கப்பட்டுள்ள மொழிப்பெயர்ப்பு பணிகள் மனதிற்கு நிறைவழிக்கக் கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிப்பெயர்ப்பிற்கு உதவுகின்றன. இவைகளைச் செம்மைப்படுத்தி முற்றிலும் சிறந்ததாக அமைப்பதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. எனவே தற்பொழுது கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிகள் வரைவு மொழிப்பெயர்ப்பிற்கு போதுமானதாகும்.

மனிதனின் படைப்பாற்றல், கற்பனைத்திறன், உளவியல் பாங்கு போன்றன செய்திகளில் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணியில் இத்தகைய கூறுகளை சேர்க்க வேண்டி உள்ளது. இதற்கு மனித நுண்ணறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன. இத்திட்டங்களில் வெற்றி பெறுகின்றபொழுது திறமையான எதிர்ப்புகளுக்கு ஏற்ற மொழிப்பெயர்ப்பு மென்பொருள்கள் உருவாக்கப்படும். அவ்வகையில் மொழிப்பெயர்ப்பிற்கு உதவக் கூடிய வரை நிலையான மொழிப்பெயர்ப்பையே கணிப்பொறி வழி கிடைக்கப்பெற இயலும்.

2.18.6. தமிழில் இயந்திர மொழிபெயர்புக்கான ஆயத்தங்கள்

தமிழ் சார் இயந்திர மொழிப்பெயர்ப்பு

தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள் பல மேற்கொள்ளப்பட்டு தமிழி நாட்டில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சிக்கு வித்திட்டன.

2.18.6.1. தமிழ் சார் இயந்திர மொழிபெயப்பு முயற்சிகள்

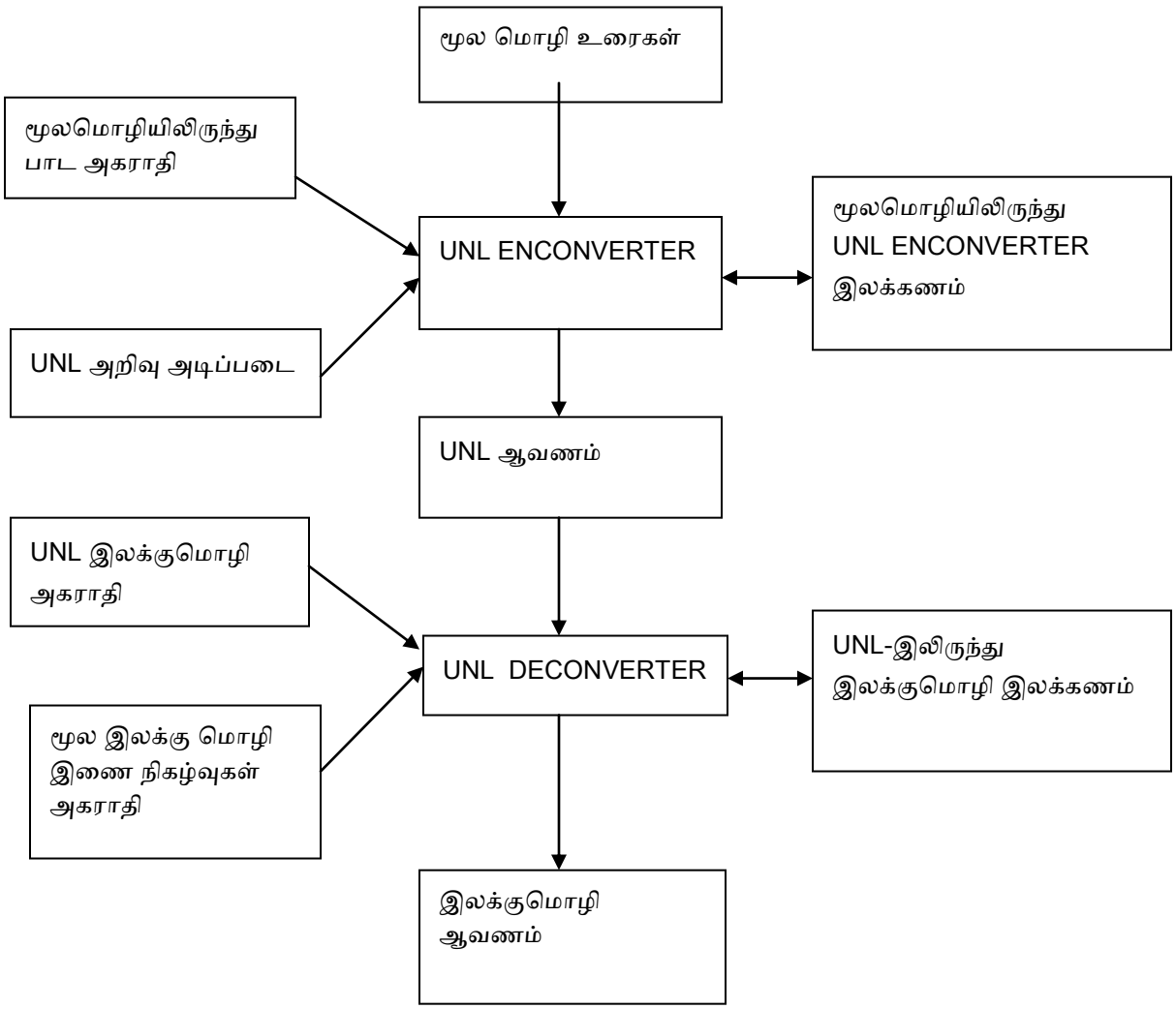
2.18.6.1.1 தமிழ் உருஷ்யன் மொழிப் பெயர்ப்புத்திட்டம்

1985-ஆம் தமிழ் பல்கலைக்கழகத்தில் (Tumtst Tamil University Machine Inaslation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. இதன் மூலம் ருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிப்பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழகம் கண்ணி அறிவியல்துறை, மொழியியல் துறை மற்றும் மொழிப் பொயர்ப்புத் துறை ஆகியவைகளின் கூட்டு முயற்சியில் உருவானதாகும்.

2.18.6.1.2 உலக வலைப்பின்னல் மொழி – தமிழ்க்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

இந்த அணுகுமுறை சென்னையில் உள்ள அண்ணா பல்கலைக்கழகத்தில் கணிப்பொறி மொழியியல் துறையில் மேற்கொள்ளப்பட்டது. (Dhanabalan & Geetha, 2004). இந்த அணுகு முறையில் இடையீட்டு மொழியின் உருப்படுத்தம் பயன்படுத்தப்படுகிறது. மூலமொழி அகராதியையும் இலக்கணத் தகவல்களையும் பயன்படுத்தி இடைமொழி உருப்படுத்தம் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிப்பெயர்ப்பு செய்ய பயன்படுத்தப்படும் இடைமொழி அமைப்பு உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)) ஆகும். இம்மொழி இடைமொழி அமைப்பாகத் திட்டமிடப்பட்டு மூலமொழியிலிருந்து இலக்கு மொழிக்கும் இலக்கு மொழியிலிருந்து மூலமொழிக்கும் தானியங்கி இயந்திர மொழிபெயர்ப்புயைச் சாத்தியமாக்குகின்றது. UNL-ஐ பயன்படுத்திய மொழிபெயர்ப்பு ஒழுங்குமுறை மூலமொழியிலிருந்து UNL-க்கு மாற Encoverter-ஐயும் UNL-இலிருந்து இலக்கு மொழிக்கு மாற Deconverter பயன்படுத்துகின்றது.

இந்த அணுகுமுறையின் திட்டவடிவம் கீழே தரப்பட்டுள்ளது.



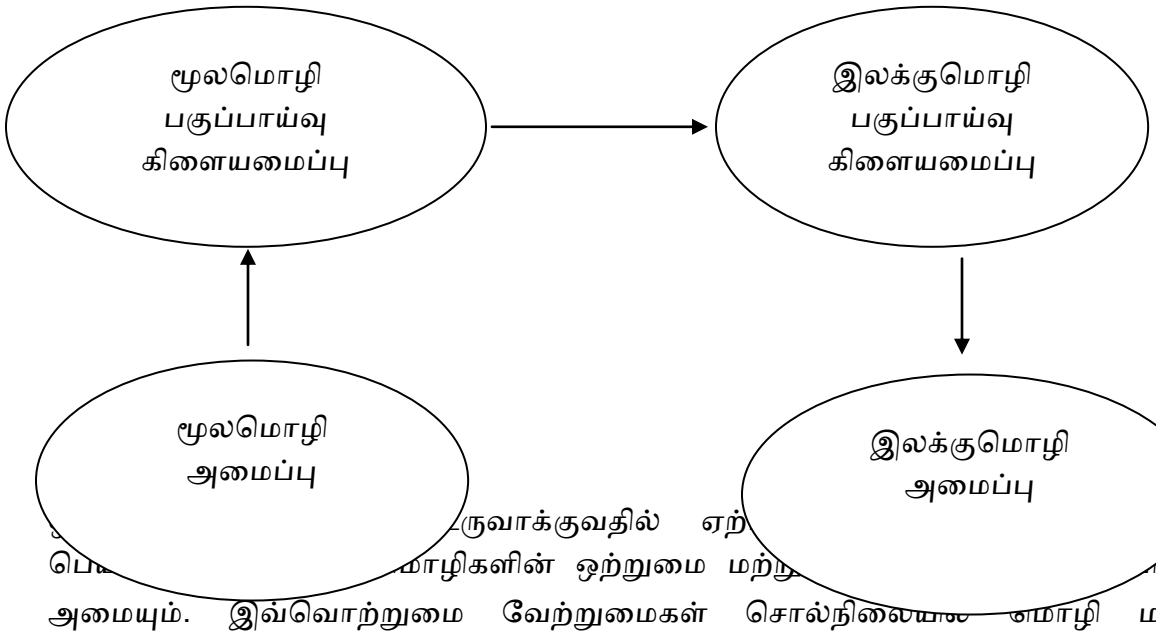
மேற்கண்டவாறு படத்தில் காட்டியவாறு UNL அமைப்பு வேலை செய்து UNL மொழியிலிருந்து தமிழுக்கு மொழிபெயர்ப்பு செய்யும்.

2.18.6.1.3 ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்க ஒரு இயந்திர மொழி பெயர்ப்புக் கருவி உருவாக்க, காமாட்சி மற்றும் ராஜேந்திரன் (Kamachi Rajendran, 2004) ஒரு திட்டத்தை உருவாக்கி செயல்படுத்த முயன்றனர். தமிழில் முதுகலை மொழியியல் பாடத்திட்டத்திற்கு வேண்டிய மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்க வேண்டிய கட்டாயம் ஏற்பட்டதால் இத்தகைய ஆய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வு மொழியியல் புத்தகங்களை ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்க உதவும் ஒரு கருவி தயாரிப்பதை நோக்கமாக கொண்டது. இக்கருவி பின்வரும் நான்கு பகுதிகளைக் கொண்டது.

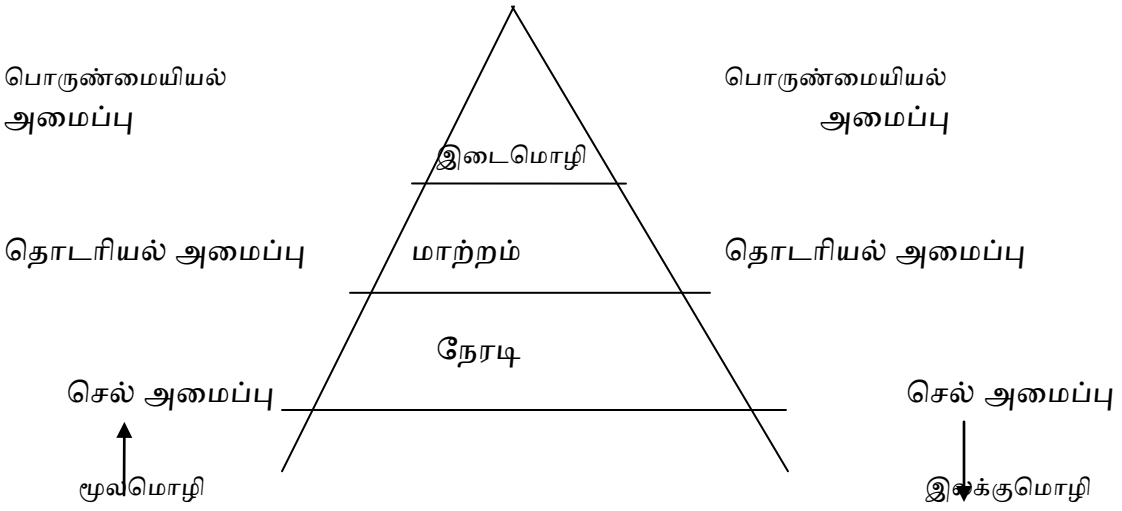
1. தமிழ் மொழி அமைப்புப் பகுத்தாய்வு
2. ஆங்கில மொழி அமைப்புப் பகுத்தாய்வு
3. ஆங்கில - தமிழ் மொழிபெயர்ப்பு அகராதி
4. ஆங்கில – தமிழ்மொழி மாற்றமைப்பு விதிகள்
5. ஆங்கில மொழியமைப்பு உருவாக்கி

இந்த மொழிப்பெயர்ப்பு கருவியின் செயல்பாட்டை விளக்கும்



உருவாக்குவதில் ஏற் மொழி
பெ மொழிகளின் ஒற்றுமை மற்று றுத்து
அமையும். இவ்வொற்றுமை வேற்றுமைகள் சொல்நிலையில் மொழி மாற்றம் செய்யப்பட வேண்டுமா? தொடர் நிலையில் மாற்றம் செய்யப்பட வேண்டுமா?

பொருண்மை நிலையில் மாற்றம் செய்யப்பட வேண்டுமா என்பதைத் தீர்மானிக்கும் பின்வரும் படம் இச்சாத்தியங்களை வெளிப்படுத்தும்.



தமிழிலிருந்து பிற மொழிகளுக்கு மற்றும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு இயந்திர உதவி மொழிபெயர்ப்பு ஒழுங்குமுறைகள் தமிழ்மொழியின் மீது பல நிலைகளிலான ஆய்வுகளை உள்ளடக்கும். ஒவ்வொரு நிலைக்கும் அதற்கான சிக்கல்கள் காணப்படும்.

2.18.6.2 தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்

2.18.6.2.1 உருபனியல் ஆய்வு (Morphological Analysis)

இது வேர் சொற்களைக் கண்டுபிடிப்பதையும் அதன் இலக்கண பண்புப் கூறுகளைக் கண்டுபிடிப்பதையும் உள்ளடக்கும். தமிழ் ஒரு திரிபு (Inflection) வளமுள்ள மொழி ஆகையால் அதன் பகுப்பாய்வு ஆங்கிலத்தின் உருபனியல் பகுப்பாய்வை விடச் சிக்கல் வாய்ந்ததாக அமையும். புணர்ச்சி விதிகள் தமிழில் உருபனியல் ஆய்வி உருவாக்குவதை கடினமாக்குகின்றது. சொற்களை உருபன்களாகப் பிரிப்பதிலும் அவற்றிற்குப் பொறுத்தமான பொருண்மை மற்றும் இலக்கணத் தகவல்கள் தருவதிலும் கவனம் தேவை. பின்வரும் எடுத்துக்காட்டு இதைக் தெளப்வுப்படுத்தும்.

தமிழ்சொல் : பணத்திற்காகத்தானே

உருபனியல் பிரிப்பு : பணம் + க்கு + ஆக + தான் + ஏ

வேர் சொல் : (பணம்)

நான்காம் வேற்றுமை உருபு (க்கு)

கொடை வேற்றுமை உருபு (ஆக)

இச்சொல் ஆங்கிலத்தில் மொழிப்பெயர்க்கப்படும் போது இது ஒரே சொல்லாக மொழி பெயர்க்கப்படாமல் கடிச வாந ளயமந அடிஇநல டிஇடல என மொழி மொழிபெயர்க்கப்பட வேண்டும்.

2.18.6.2.2 சொல் வகைப்பாட்டு அடையாளப்படுத்தல் (Parts of Speech Tagging)

சொல் வகைப்பாட்டு அடையாளப்படுத்தி வாக்கியத்திற்கும், சொற்களுக்கும் சொல்வகைப்பாடு தரவேண்டும். இது வாக்கியத்தின் அமைப்பைப் புரிந்துக்கொள்ளவும் வாக்கியத்தின் அமைப்பை உருவாக்கவும் முக்கியமானதாகும்.

தமிழ் வாக்கியம் : அவன் படி ஏறினான்

He staircase Climbed

மேற்கண்ட வாக்கியம் பின்வருமாறு அடையாளப்படுத்தப்படும்.

அவன் < PRO >

படி < N+PRO >

ஏறினான் < V+PST+3SM >

இதில்,

PRO ⇒ Pronoun ‘ மாற்றுப் பெயர் ’

N ⇒ Noun ‘ பெயர் ’

NOM ⇒ Nominative ‘ எழுவாய்ப் பெயர் ’

V ⇒ Verb ‘ வினை ’

PST ⇒ Past tense ‘ இறந்த காலம் ’

3SM ⇒ 3 Person Singular, Masculine ‘ படர்க்கை ஆண்பால் ’

இந்த வேர்ச்சொல்லான ‘ படி’ ‘ staircase’ என்ற பொருளையோ ‘ read’ என்ற பொருளையோ தரலாம். சோல்வாகைப் பாட்டு அடையாளப்படுத்தி சொல் மற்றும் தகவலைப் பயன்படுத்தி இந்த பொருள் மயக்கத்தைத் தீர்க்கும். இவ்வாறு ‘ படி’ என்பதற்குப் பெயர் சொல்வகைப் பாட்டு அடையாளம் தரப்படும்.

2.18.6.2.3 தொடரியல் ஆய்வு (Syntax analysis)

தொடரியல் ஆய்வின் போது பகுத்தாப்படுகையில் வாக்கிய அமைப்பு தெரிந்து கொள்ளப்படும். பகுத்தாய்வதால் எச்சத்தொடர் எல்லைகள், முன்னுருபு இணைப்பு அடைகள் என்பன தெரிந்துக்கொள்ளப்பட்டு அவற்றின் பொருத்தமான உறுப்புகளுடன் தொடர்புப் படுத்தப்படும். தமிழ் ஒரு சொல் சுதந்திரமான மொழியாகும். இது தொடரியலில் பகுத்தாய்தலைச் சிக்கலாகும். பெயருடன் இணைந்து வரும் வேற்றுமை உருபுகள் வாக்கியத்தில் இவ்வுறுப்புகளின் பங்களிப்பைக் கண்டு பிடிக்கின்றன. ஆனால் வேற்றுமை உருபு இல்லாது இருப்பு அமைப்பு சார் பொருள் மயக்கத்திற்கு வழிவகுக்கும். எடுத்துக்காட்டாக,

தமிழ் வாக்கியம் : ‘ அவள் அக்காள் வீட்டிற்குச் சென்றாள்’

‘ She / he Sister house – dative go+pst 2sf’

என்பதற்கு இரண்டு பகுத்துக் குறிப்புகள் கிடைக்கும் :

(அவள்) NP (அக்காள் வீட்டிற்கு) NP சென்றாள் V

(அவள் அக்காள்) NP (vii TT : RKU) NP சென்றாள் V

2.18.6.2.4 பொருண்மையியல் ஆய்வு (Semantic Analysis)

பல அர்த்தங்கள் உள்ள சொற்களுக்கு ஒரு குறிப்பிட்ட சூழலுக்குத் தகுந்தவாறு சரியான அர்த்தம் கண்டுபிடிக்கப்பட வேண்டும். இலக்குமொழி நிகரன் மயக்கம் நீக்கப்பட அர்த்தம் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும்.

தமிழ் வாக்கியம் 1 : அவன் படி ஏறினான்

Climb + pst + 3sm

தமிழ் வாக்கியம் 2 : அவன் படியால் அளந்தான்

He Vessltins Measuvet Ps ++3sm

மேற்கூறிய வாக்கியங்களில் படி என்பதற்கு இரண்டு அர்த்தங்கள் இருக்கின்றன.

1. Staircase
2. A vessel used to measure rice and whet.

முதல் வாக்கியத்தில் ஏறினான் என்ற சொல்லுடன் இணைந்து வருகையைப் பயன்படுத்தி படி என்பதன் அர்த்தம் ‘ Staircase’ என்பதுதான் என்றும் இரண்டாம் வாக்கியத்தில் அளந்தான் என்ற சொல்லுடன் சேர்ந்து வருகையால் படி என்பதன் அர்த்தம் ‘ vessel’ என்பதுதான் என்றும் நிர்ணயிக்கப்படுகின்றது.

2.18.6.2.5 தொடரியல் சார் மாற்றம் (Syntactic Transfer)

மொழிபெயர்ப்பு செல்பாட்டில் மாற்றல் அணுகுமுறை (transfer approach) பயன்படுத்தப்படும் போது தொடரியல் மாற்றப்பகுதி தேவைப்படும். ஊள்ளீடு செய்யப்படும் வாக்கியத்தின் அமைப்பு இலக்குமொழியின் பொருத்தமான அமைப்பிற்குத் தகுந்தவாறு மாற்றப்படுகிறது.

தமிழ் வாக்கியம் : ராமன் புத்தகம் படித்தான்

Raman book read

அமைப்பு: Noun Noun Verb

மேற்கண்ட வாக்கியங்களுக்கு நிகரான வாக்கியங்கள் கீழ்வருவன.

ஆங்கில வாக்கியம் : Raman read a book
அமைப்பு :Noun Verb Noun

இந்த மொழிபெயர்ப்பில் Noun Noun Verb என்ற தமிழ் வாக்கியம் Noun Verb Noun என்று ஆங்கிலத்திற்கு நிகராக மாற்றப்படுகிறது. இந்த எடுத்துக்காட்டு மிக எளிமையானதாகும். ஆனால் இச்செயல்பாடு கலவை வாக்கியங்களுக்கு கிடையில் கடினமானதாகும்.

2 .18.6.2.6 அகராதி பொருத்தம் (Dictionary Mapping)

இருமொழி அகராதியைப் பயன்படுத்தி மூலமொழி வேர் சொற்கள் இலக்குமொழியில் அதற்கு நிகரான சொற்களுடன் பொருத்தம் செய்யப்படும். ஒரு மூலமொழி வேர்ச்சொல்லுக்கு இலக்கு மொழியில் பல சொற்கள் கொண்ட தொடர்கள் நிகரன்களாக வரலாம். சூழல் அடிப்படையில் பொருத்தமான சொல் சூழல் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும். வேர்ச் சொல்லின் இலக்கண பண்புக் கூறுகளின் இலக்கணம் அகராதியில் அடங்கி இருக்கும்.

2.18.6.2.7 உருபனியல் உருவாக்கம் (Morphdogical Generation)

இங்கு சொல்லின் திரிபு வடிவம் வேர் மற்றும் அதன் இலக்கண பண்புக் கூறுகளால் உருவாக்கப்படும்.

தமிழ் வேர்ச்சொல் : ‘ படி’

இலக்கண பண்புக் கூறுகள் : ‘ Past tense’ + 3rd person, Singular, Mascuine’

சொல் வடிவு : ‘ படித்தான் ’

2.18.6.3 தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்

2.8.3.1 கருவிகள் (Tools)

2.18.6.3.1.1 உருபனியல் ஆய்வி (Morphoogical Analyzer)

95 விழுக்காட்டிற்கு மேல் முழுமையுடன் தொழில் நுட்பத்தின் தற்போதைய நிலையைப் பயன்படுத்தி ஒரு திறமையான உருபனியல் ஆய்வி தேவை. தற்போதைய உருபனியல் ஆய்விகள் மேற்கோள்வாய்வாய்பாட்டு (Paradim approach) அணுகுமுறையையும் முற்றுநிலைத் தானியங்கியையும் (Finite State Automata) பயன்படுத்துகின்றனர்.

2.18.6.3.1.2 அடையாளப்படுத்தி

வாக்கியங்களை அடையாளப்படுத்த முழுதானியக்க, பகுதிதானியக்க, விதி அடிப்படையிலான அல்லது புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தித் தேவை. இவ்வகையான அடையாளப்படுத்தல் இயந்திரமொழி பெயர்ப்புக்கு மட்டுமல்லாமல் தகவல் பிரித்தெடுப்பு, தகவல் மீட்பு உரை சுருக்கம் மற்றும் பிற பல்வேறுபட்ட பயன்பாடுகளிலும் பயன்படுகிறது.

2.18.6.3.1.3 பகுத்துக் குறிப்பான் (Parser)

ஒரு பகுத்துக் குறிப்பான் சார்பு இலக்கணம் (dependency Gramman) தொடரமைப்பு இலக்கணம் (Phrase Structure Grammer), கிளை இணைப்பு இலக்கணம் (Tree

Adjoining Grammar) போன்ற ஏதாவது இலக்கண வடிவமைப்புகள் அடிப்படையில் வாக்கியங்களுக்குப் பகுப்புக் கிளை தரும். ஒரு பகுத்துக் குறிப்பான் தமிழ்மொழிப் பகுப்பாய்விற்கு மிகத் தேவை. இது விதிகள் மற்றும் பகுத்துக் குறிக்கும் இயந்திரத்தை உருவாக்குதலையும் உள்ளடக்கும். சில இயற்கை மொழி ஆய்வுப் பயன்பாடுகள் முழு பகுத்துக்குறித்தலை வேண்டாது, அவைகளுக்குப் பகுதி பகுத்துக் குறித்தல் போதுமானது. பெயர்த்தொடர் கண்டுபிடிப்பு, எச்சத்தொடர் கண்டுபிடிப்பு என்பன இந்தப் பயன்பாடுகளுக்கு உதவும்.

2.18.6.3.1.5 பொருண்மையியல் ஆய்வி (Semantic Analyser)

சொல் அர்த்த மயக்கம் நீக்கும் கருவியை உருவாக்கும் வேலை அதன் தொடக்க நிலையில் தான் இருக்கிறது. இந்த வேலை கூடுதலான கவனக் குவிப்பையும் கடினமான முயற்சிகளையும் வேண்டும். மேலும் இது சொற்கள் இணைந்து வரும் தகவலைச் சேகரிக்க வேண்டி பெரிய அடையாளப்படுத்தப்பட்ட விரிதரவை வேண்டும்.

2.18.6.3.1.6 உருபனியல் உருவாக்கி (Morphological Generator)

நம்மிடம் உருபனியல் ஆய்வுக்கு நன்றாக அமைக்கப்பட்ட முற்றுநிலை இயந்திரம் இருந்தால் அதை உருபனியல் உருவாக்குதலுக்குச் சிறிது மாற்றமின்ற அல்லது சிறிது மாற்றத்துடன் அதே இயந்திரத்தைப் பயன்படுத்தலாம். இவ்வாறு உருவாக்கப்பட்ட உருபனியல் உருவாக்கிகள் திருப்தியான விளைவுகளைக் காட்டுகின்றன.

2.18.6.3.2 மூலவளங்கள் அகராதி (Resources)

2.18.6.3.2.1 அகராதி (Dictionary)

பல அகராதிகள் தமிழில் இருந்தாலும் அவையெல்லாம் மின்வடிவில் இல்லை. இந்த அகராதிகள் இயந்திரம் படிக்கவியலும் வடிவத்தில் இருந்தால்தான் அவற்றைக் கணினி செயல்பாட்டிற்குப் பயன்படுத்த இயலும்.

2.18.6.3.2.2 மொழி கடந்த அகராதி (Cross -lingual Dictionary)

இது சாதாரண மொழிகள் கடந்த அகராதியைக் குறிப்பிடவில்லை. இயந்திர மொழிபெயர்ப்புக்கு நமக்கு தேவையான அகராதி மொழி கடந்த வேர் அகராதிகளாகும். தனிப்பட்ட மனிதர்களும் நிறுவனங்களும் அவர்களுடைய பயன்பாட்டிற்காக இருமொழி அகராதிகளை உருவாக்குகின்றனர். இவை இந்த ஆய்வு களத்தில் சூடுபடும் ஆய்வாரள்களுக்கிடையில் பங்கிடப்படவேண்டும்.

2.18.6.3.2.3 தரவுத்தொகுதி (Corpus)

ஒரு மொழி தரவுத்தொகுதி அம்மொழியின் நடை, சொல், பயன்பாடு போன்றவற்றின் உருப்படுத்தம் ஆகும். இது மேலும் உருபனியல் பகுப்பாவிகள், பகுத்துக் குறிப்பான்கள் மற்றும் உருவாக்கிகள் என்பனவற்றை மதிப்பீடு செய்யும் பரிசோதனைக் கருவியாகச் செயல்படும். இந்திய மொழிகளின் மைய நிறுவனம் சுமார் 15 மொழிகளுக்கு மூன்று மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதிகளை உருவாக்கியுள்ளது. மேலும் சில தரவுத்தொகுதிகளை அடையாளப்படுத்தி உள்ளது. இருக்கின்ற வசதிகளை வைத்துக்கொண்டு பல்வேறுபட்ட மொழிகளுக்கு விரிதரவைச் சேகரிக்க வேண்டி பொதுவான திட்டம் ஊக்கப்படுத்தப்பட வேண்டும். புத்தகங்கள், நாளிதழ்கள், கால இதழ்கள் போன்றவற்றஇ தவிர இணையதளங்களைப் பயன்படுத்தி தரவுத்தொகுதி சேகரிக்கப்பட்ட வேண்டும்.

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட விரிதரவில் (POS Tagged Corpus) உள்ள எல்லா வாக்கியங்களும் சொல்வகைப்பாட்டிற்கு வேண்டி அடையாளப்படுத்தப்பட்டிருக்க வேண்டும். இந்த அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, புள்ளியியல் அடிப்படையிலான சொல் வகைப்பாட்டு அடையாளப்படுத்தி உருவாக்குவதற்குத் துணைபுரியும். இம்மூலவளத்தின் மீது பல்வேறு வகையான ஆய்வுகளினால் ஒரு மொழிக்குக் கணினி சார் இலக்கணம் உருவாக்க இயலும்.

இணை தரவுத்தொகுதிகள் (Parellel Corpora)

இணைதரவுத்தொகுதி இருமொழிகளுக்கு மாற்றமைவு விதிகளை (transfer rules) உருவாக்க முக்கியமான மூலவளமாகும். இது புள்ளியியல் அல்லது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறைகள் உருவாக்கத் துணைபுரியும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கு முன் வரிசையாக்க விதிகள் இணைவிரிதரவிகளின் பத்திகள், வாக்கியங்கள் சொற்கள் இவற்றை வரிசையாக்கம் செய்ய வரிசையாக்கக் கருவிகள் தேவைகள். சுப்ல வெளிப்யீட்டாளர்களிடமும் அரசு நிறுவனங்களிலும் இம்மாதிரியான இணைதரவுத்தொகுதிகள் இருக்கின்றன. இந்த இணைதரவுத்தொகுதிகள் விளைவாக்கம் உள்ள ஒழுங்குமுறைகளை உருவாக்க வேண்டி ஆராய்ச்சியாளர்களுக்குப் பயன்படுத்தத் தரப்படவேண்டும்.

2.18.6.3.2.4 மாற்றமைவு இலக்கணம்

மாற்றமைவு இலக்கணம் (transfer grammar)இருமொழிகளின் தொடரியல் அமைப்புக்கு இடையில் உள்ள முரண்வேறுபாடுகளை அறிந்து கொள்கின்றது. மாற்றமைவு அணுகுமுறைகளை உருவாக்க இம்மாதிரியான மாற்றமைவுப் பகுதியை உருவாக்குவது கட்டாயமாகும். மொழிகளைப் பற்றிய ஆழ்ந்த அறிவுள்ள மொழியியலார் இம்மாதிரியான மூல வளத்தை உருவாக்கலாம்.

2.18.6.3.2.5 இணைவமைதி அகராதி

இணைவமைதி அகராதி (Collocation dictionary) மொழியின் மயக்கமான சொற்களின் முறையான சேர்க்கைகளைக் கொண்டிருக்கும். அம்மாதிரியான அகராதி அர்த்த மயக்க நீக்கிற்கு முக்கியமான பகுதியாகும். தமிழுக்கு அம்மாதிரியான மூலவளத்தைச் சில ஆய்வாளர் குழுமங்கள் செய்து வருகின்றன.

2.18.6.9 நிறைவுரை

இவ்வியலில் கணிப்பொறி வழி இயந்திர மொழிபெயர்ப்பின் தேவை, கணீபொறி வழி மொழிபெயர்ப்பின் வகைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும், இந்தியாவில் இயந்திர மொழி பெயர்ப்பின் வளர்ச்சி, தமிழ்ச்சார்ந்த இயந்திர மொழிபெயர்ப்புகள் என்ற தலைப்புகளில் விளக்கங்கள் தரப்பட்டுள்ளன.

இயல் 3

முடிவுரை

மேற்கண்ட பகுதிகள் மொழி ஆய்வில் கணிப்பொறியின் முக்கியத்துவத்தையும் தமிழில் இயற்கை மொழி ஆய்வின் பரிமாணத்தையும் நமக்கு வெளிப்படுத்துகின்றன. கணிப்பொறி மொழியியல் ஆய்வுகளுக்கு ஒரு வரப்பிரசாதமாய் மாறியுள்ளது. மொழியியல் ஆய்வுகள் கணிப்பொறி அறிவியலுக்குத் தேவை என்ற நிலையுடன் கணிப்பொறி மொழியியல் ஆய்வுகளுக்கு மிகத் தேவை என்ற நிலையும் இன்றைய கால கட்டத்தில் நிலவுகின்றது. கணிப்பொறி மொழியின் அமைப்பை ஒலியனியல்,

உருபனியல், தொடரியல், பொருண்மையியல் என்ற மட்டங்களில் ஆய்வதற்கு உதவுவதுடன் மொழிக்கருவிகள் உருவாக்குவதற்கும் இன்றியமையாத சாதனமாய் பயன்படுகின்றது. அட்டைகளில் பதிவுகளையும் பொருள்களையும் எழுதும் காலம் போய் அகராதி தயாரிக்கும் மென்பொருள்களைப் பயன்படத்தி அகராதிகள் தயாரிக்கும் காலம் நிலவுகின்றது. கணிப்பொறியின் வரவால் அகராதி தயாரிப்பது ஒரு தண்டனையாக இல்லாமல் இனிய அனுபவமாக மாறிவிட்டது. ஆசிரியரின் இடையீட்டால் உற்சாகமில்லாத செயல்பாடாக இருந்த மொழி கற்றல் கணிப்பொறியின் உதவியால் நினைத்தால் இனிக்கும் அனுபவமாக மாறியுள்ளது. கற்றல் மென்பொருள்களைப் பயன்படுத்தி குழந்தைகள் உற்சாகமாக விளையாடிக்கொண்டே மொழியைக் கற்கின்றனர். அகரவரிசை அட்டவணை படுத்துவதற்கும் சொல் அடைவுகள் தயாரிப்பதற்கும் அட்டைகளைப் பயன்படுத்துவதைக் கணிப்பொறி இல்லாமல் செய்துவிட்டது. உரையிலுள்ளதை பேச்சாகவும் பேச்சை உரையாகவும் மாற்றும் மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. எழுத்துணரியின் வரவால் கணிப்பொறியில் தட்டச்சுசெய்து உரைகளை உள்ளீடு செய்யவேண்டும் என்ற நிலை மாறியுள்ளது. கடுமையான வேலையாக இருந்த மொழிபெயர்ப்பு கணிப்பொறியின் பயன்பாட்டால் எளிய வேலையாக மாறியுள்ளது. எழுத்துரையை மட்டுமன்றி பேச்சுரையையும் உடன்தானே மொழிபெயர்க்கும் VERB MOBILE போன்ற கருவிகள் உருவாக்கபட்டு பயன்பாட்டிலுள்ளன. இனி வரும் காலத்தில் செல்போனில் ஒருவர் ஆங்கிலத்தில் பேசினால் மற்றவர் அதைத் தமிழில் கேட்கவியலும். இதுவரை முன்னேற்றத்தின் பரிமாணத்தையும் வேகத்தையும் கணிக்க முடிந்தது. கணிப்பொறியின் பயன்பாட்டால் முன்னேற்றத்தின் பரிமாணத்தையோ வேகத்தையோ கணிக்கவியலாத சூழல் உருவாகியுள்ளது. நாளைய உலகம் கடவுள் கையில் என்று கூறவியலாமல் கணிப்பொறி கையில் என்று கூறத்தோன்றுகின்றது.

துணைநூல்கள்

சுப்பையா பிள்ளை, 2000. இயற்கை மொழி ஆய்வு. உலகத்தமிழாராய்ச்சி நிறுவனம், சென்னை.

செவ்வேள் கபிலன். 1994. கணிப்பொறிவழி தமிழ் வினைகளின் பகுப்பாய்வு. புத்தாக்க மொழியியியல் கழகம், சென்னை

பாஸ்கரன், ச. 2004. தமிழில் கணிப்பொறியியல், கணிப்பொறியில் தமிழ். உமா பதிப்பகம், தஞ்சாவூர்.

Anbarasan, N. 2001. “Tamil Unicode-What do we need?”, paper read in National Seminar on Computational Linguistics and Dravidian Languages. CAS in Linguistics, Annamalai University, Annamalai.

Anandan, P, Rajani Parthasarathy, Geetha, T.V. 2001. “Morphological Generator for Tamil”, in Proceedings of the Tamil Internet 2001 Conference.

Annamalai, E. “Corpora Development in Indian Languages”, in Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

Arulmozhi, P., Sobha, L., Kumara Shanmugam, B. 2004. Parts of Speech Tagger for Tamil, Symposium on Indian Morphology, Phonology & Language Egnineering, Indian Institute of Technology, Kharagpur.

Arulmozhi, P and Sobha, L. 2006. A Hybrid POS tagger for a Relatively Free Word Order Language. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, pages 79-85.

Arulmozi, S. 1998. Aspects of Inflectional Morphology–A Computational Approach. Ph.D dissertation submitted to University of Hyderabad.

Arulmozi, S and M.C. Kesava Murty. 2006. Verb Sense Disambiguation in Telugu-Tamil MT. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages pages 107-111.

Balakrishnan, R. 2002. Morphology and Tamil Computing. Paper read in International Seminar on Tamil Computing, February 27, 28, 2002, Madras University.

Baskaran, S. “User Interface with Computers in Tamil-an Overview [*kaNippoRi uraiyaakam*]”, in Proceedings of the Third Tamil Nadu Science Congress, Pondicherry.

Baskaran, S. “Zip’s Law: A perspective study in Tamil”, in Proceedings of Eighth International Conference of Tamil Studies.

Baskaran, S. Report on Standardization of Tamil Key Board, Submitted to the Committee formed for Standardization of Tamil Key board by Tamilnadu Government.

Baskaran, S. “Tamil Computing”, *tamizh vaLarcci*.

Baskaran, S. “Evolution of Tamil computing”, Souvenir, National Symposium on Current Trends in Computer Applications.

Baskaran, S. “Experiment in Implementation of Finite-State-Machine Parser for Tamil Sentences”, in Computer Society of India Communications, Bombay.

Baskaran, S. “Compute as a Language Research Tools,” in Proceedings of the Second Tamilnadu Congress, Madras.

Baskaran, S. 2002. Word Sense Disambiguation of Tamil. MS dissertation submitted to Anna University, Chennai.

Baskaran, S. 2003. “Word Sense Disambiguation in Tamil”, International Journal of Dravidian Linguistics.

Baskaran, S and Vaidehi, V. 2004. “Collocation Based Word Sense Disambiguation using Clustering for Tamil”, International Journal of Dravidian Linguistics 33.1: 13-28.

Baskaran, S. Vijay-Shanker, K. 2003. “Influence of Morphology in Word Sense Disambiguation for Tamil”, in Rajeev Sangal et al (eds.) Recent Advances in Natural Language Processing. Mysore: Central Institute of Indian Languages.

Brochure on ‘Language Technology Products’ of the Resource Centre for Indian Language Technology Solutions-Tamil, Chennai.

Chellamuthu, K.C. et al. Tamil University Machine Translation System (TUMTS), Thanjavur: Tamil University.

Chellamuthu, K.C. 2002. ‘Russian to Tamil Machine Translation System at Tamil University,’ in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, 74-83.

Chithiraputhiran, H. Tamil Lexical Resource: An electronic Data Base. Paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Deivasundaram, N. and Gopal, A. 2003. 'Computational Morphology of Tamil' In B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 406-410.

Dhurai Pandi, 2002. "Morphological Generator and Parsing Engine for Tamil Verb Forms (abstract)", in Kalyanasundaram (ed) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 59.

Durai Pandi. 2002. "English-Tamil Machine Translation System", in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 86.

Durai Pandi, 2002. "Morphophonemic Rules for Tamil Computing", in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 98.

Elangovan, A. "Optimisation Techniques in Unicode Tamil Font Development", in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 30.

Final Report: Development of Corpora of Texts of Indian Languages in Machine Readable form: Part II (Tamil, Telugu, Kannada, Malayalam), TDIL-Corpora Group, CIIL, Mysore January 1995.

Francis Ekka and Ganesan, M. 1994. "Issues on Standard ISCII Codes and Inscript Keyboard", Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

'The Unicode Standard: A Technical Introduction' 8/05/2001.

Ganesan, M. 1994. "Functions of the Morphological Analyser Developed at CIIL, Mysore", in Harikumar Basi (ed.) Automatic Automatic Translation (seminar proceedings), Thiruvananthapuram: ISDL.

Ganesan, M. 1994. "'A Scheme for Grammatical Tagging of Corpora of Indian Languages", in B.B. Rajpurahit (ed.) Technology and Languages, Mysore:CIIL

Ganesan, M. 1996. "Relevance on Indian Grammatical Theories for Automatic Translation: chances and challenges", paper presented in National Seminar of Dravidian Linguists Association, Dravidian University, Kuppam, June 1996.

Ganesan, M. 2000. "Compilation of Electronic Dictionary for Tamil", Annamalai University, Annamalai Nagar. www.bhasshaindia.cjb.net/G7ganesan.pdf

Ganesan, M. 2003. “Computational Morphology of Tamil”, in B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, pages 399-405.

Ganesan, M and Francis Ekka. 1994. “Morphological analyzer for Indian Languages”, Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, Newdelhi: BPB Publication.

Hewavitharana, S. 2002. ‘A Two Stage Classification Approach to Tamil Handwriting Recognition’ in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 118-124.

Information Technology Department. ‘Information Technology – Standardization of Tamil Key Board and Encoding of Tamil. Glyphs – Recommendations of the Sub-committee on Tamil in Information Technology – Accepted – orders – Issued. (G.O. Ms.NO.17 dated” 13 June 1999)

Jayaram, B.D. “Development of Corpora in Indian Languages”, paper presented in the Seminar in the use of Computers in Indian Languages, held in CIIL, Mysore, August 1992.

Kalyanasundaram, K.A. “Comparison of transliteration schemes and key mapping of Tamil fonts”, http://www.Geocities.com/Athens/5180/translit.html 19/12/2000

Kamakshi, S. 2002. “Preliminaries for Digitizing the Personal Pronouns in English into Tamil (Distribution Sensitive Machine Translation Aid–DSMTA)–A Demo Paper”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 87-97.

Kamakshi, S. 2004. “Machine Recognition and Translation of ‘~ing’ words in English into Tamil through bilingual Machine Tractable Dictionary”, paper presented in International Conference on Indian Lexicography 28-30 January 2004, CASL, Annamalai University, Annamalainagar.

Kamakshi, S. and Rajendran, S. 2004. Preliminaries to the preparation of a machine aid to translate Linguistics Texts written in English into Tamil. DLA Publications, Thiruvanathapuram.

Kasirao, V. “Impact of Information Technology on Information Management Sevices with Special Reference to Tamil Computing: A Study”, paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Krishnamoorthy, V. “OCR Software for Printed Tamil Text”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 99-101.

Kumara Shanmugam, B. 2001. “Parse Representation of Tamil Syntax”,

Kumara Shanmugam, B. 2002. “Machine Translation as related to Tamil”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 84-85.

Kumara Shanmugam, B. 2004. Syntactic Parser for Tamil. M.S. (in NLP) dissertation submitted to Anna University, Chennai.

Lyons, J.1977. Semantics, volume 1, New York: Cambridge University Press.

Michael S. Kapalan. 2002. “Unicode and Tamil”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 1.

Muthu Nedumaran. 2002.”Glyph Choices and Techniques for Building Unicode Based Tamil Fonts”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 31.

Natanasabapathy, S. 2002. “Study of Segmental Duration Tami sounds”, paper presented in International Seminar on Tamil Computing, 27-28 February and 1 March, 2002, University of Madras, Chennai.

Nida, E.A. 1975.a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.

Nida, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton.

Navaneethan, P et al. 2004. “Multilingual Word Processor with Online Noun Translator”, paper read in National Seminar on Cross Language Communication Tools (Machine Translation) for Indian Languages, February 25-27, 2004.

Nayeemulla Khan, A, Suryakanth, V, Gangashetty, & Yegnanarayana, B. 2003. “Syllabic Properties of Thee Indian Languages: Implications for Speech Recognition and Language Identification”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Ponnavaiko, M. 2002. “An Investigation on Unicode Standards for Tamil”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 29.

Prasanna Venkatesan, P and Chitrakekha T and Kuppuswami P. 2002. “PONN: A Tamil Operating Environment”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers.

Prathibha, P., Ramakrishana, A.G. 2002. “Web-enabled Speech Synthesizer for Tamil”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, pages 134-140.

Prathibha, P., Ramakrishana, A.G. and Muralishankar, R. 2002. “Thirukural-II – A Text-to-speech Synthesis System”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, 126-133.

Rajan, K and Ganesan, M. 2002. “Tools for Corpus Analysis”, paper presented in International Conference on South Asian Linguistics (ICOSAL-4), 3-5 December, 2002, CASL, Annamalai University, Annamalainagar.

Rajendran, S. 2002. “Preliminaries to the Preparation of wordnet for Tamil”, Language in India2:1, www.languageinindia.com

Rajendran, S. 2003. “Dravidian WordNet: a proposal”, in R.M. Sundaram, et al (eds) Facets of Language, Thanjavur

Rajendran, S. 2003. “Creating Generative Lexicon form Dictionaries: Tamil Experience”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Rajendran, S. 2003. “Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages”, Language in India 3:1, www.languageinindia.com

Rajendran, S. 2006. “Shallow Parsing in Tamil: the state of art”, Language in India 6:7, www.languageinindia.com

Rajendran, S. 2006. “Languge Technology in Tamil”, Language in India 6:8, www.languageinindia.com

Rajendran S, Arulmozi S, Ramesh Kumar S, & Viswanathan S. 2003. “Computational Morpohology of Verbal Complex”, in B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 376-398.

Rajendran, s. and Baskaran, S. 2002. "Preparation of Electronic Thesaurus for Tamil", in Proceedings of the International Conference on Natural Language Processing. Mumbai: NCST.

Rajendran, S. and Kamakshi, S. 2002. "Preliminaries to the preparation of a machine translation aid to translate Linguistics Texts in English into Tamil", Paper read in UGC-SAP National Seminar 'On Translation' 7th-9th March 2002, CAS in Linguistics, Annamalai University.

Ramakrishnan, A.G. "Issues in Standardization for Text to Speech in Tamil", Tamilnet 2001, Kulalumpur, Malasia.

Ramaswamy, V. 2000. Morphological Generator for Tamil. Unpublished M.Phil. dissertation. University of Hyderabad.

Ranganathan, V. 1997. "A Lexical Phonology Approach to Tamil Words by Computer", International Journal of Dravidian Linguistics 26:1.57-70.

Shanmugan, C. 2001. "Computer Analysis of Simple Sentences in Tamil", Paper read in UGC-SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, Annamalainagar

Shanmugan, C. 2002. "Grammar and Parser: A Program for Syntactic Parsing in Tamil", International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Shanmugan, C. "Minimalist Program for Tamil Parsing",

Sobha, L and Vijay Sundar Ram. 2006. "Noun Phrase Chunker for Tamil Language", in Proceedings of the First National Symposium on Modeling and Shallow Paring of Indian Languages, pages 194-198.

Suresh, R.M., Arumugan, S., and Aravanan, K.P. 2000. "Recognition of Handwritten Tamil characters using fuzzy classificatory approach", in Proceedings of the Tamil Internet 2000 Conference, Singapore.

Vaishnavi Ramaswamy. 2000. A Morphological Generator for Tamil. M.Phil Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi Ramaswamy. 2003. A Morphological Analyzer for Tamil. Ph.D Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi, Ramaswamy. 2003. “Parsing in AMPLE, KIMMO & PERL: Nouns in Tamil”, Paper presented in

Vasudevan, V. 2002. “Character Recognition Techniques–A Demo Program”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Vasu Renganathan. 2002. “Interactive Approach to Development of English-Tamil Machine Translation System on Web”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 68-73.

Viswanathan, S., Rameshkumar, S. Kumara Shanmugam, B. Arulmozhi. S. & Vijay Shnakar, K.2003. “A Tamil Morphological Analyzer”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Winston Cruz, S. 2002. Parsing and Generation of Tamil Verbs in GSMorph. M.Phil. dissertation submitted to the University of Hyderabad.

Yagnanarayana, B. et al. 1992. “Text –to-speech system for Indian Languages”, presented at the Workshop on Computer Applications in Indian Languages held at CIIL, Mysore, August 19-21

Yegnanarayanan B, Rajendran, S, Ramachandaran V.R., Madhukumar, A.S. “Significance of Knowledge Sources for a Text-to-speech System for Indian Languages”, in Sadhana, Academic Proceedings in Engineering Sciences of India.

Yegnanarayanan, B. 1994. “Speech synthesis by Machine”, in B.B. Rajpurohit (ed.) Technology and Languages. Mysore:CIIL.